Use of auxiliary information in survey sampling at the sampling stage and the estimation stage

Éric Lesage

ANNÉE 2013

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et applications

Ecole doctorale Matisse

présentée par

Éric LESAGE

Préparée à l'Institut de recherche mathématique de Rennes
(IRMAR, UMR 6625)
à l'École nationale de la statistique et de l'analyse de l'information
(ENSAI)

Utilisation d'information auxiliaire en théorie des sondages à l'étape de l'échantillonnage et à l'étape de l'estimation

Thèse soutenue à l'ENSAI
le 31 octobre 2013

devant le jury composé de :

Yves BERGER
Professeur, University of Southampton / *rapporteur*

Hervé CARDOT
Professeur, Université de Bourgogne / *rapporteur*

Éric GAUTIER
Professeur à l'ENSAE-ParisTech / *examinateur*

Valérie MONBET
Professeur, Université de Rennes 1 / *examinateur*

François COQUET
Professeur, ENSAI / *directeur de thèse*

Jean-Claude DEVILLE
INSEE / *co-directeur de thèse*

Utilisation d'information auxiliaire en théorie des sondages à l'étape de l'échantillonnage et à l'étape de l'estimation.

Éric LESAGE
ENSAI

Thèse de Doctorat

31 octobre 2013

Remerciements

Je remercie très chaleureusement mes deux directeurs de thèse François Coquet et Jean-Claude Deville. Je dois reconnaître le rôle déterminant qu'a joué Jean-Claude Deville, par ses enseignements et ses travaux de recherche, dans mon orientation vers les sondages. Je suis très reconnaissant envers François Coquet pour le soutien sans faille qu'il m'a apporté tout au long de ma thèse.

Je remercie très sincèrement mes deux rapporteurs, Hervé Cardot et Yves Berger, pour le temps précieux qu'ils m'ont accordé. Je remercie également Éric Gautier et Valérie Monbet pour leur participation au jury.

Je remercie l'Insee, et tout particulièrement Alain Charraud et Pascal Chevalier, d'avoir soutenu mon projet de recherche et de m'avoir permis de rejoindre le laboratoire de statistique d'enquête du Crest-Ensai en juin 2010.

Je remercie l'ensemble des membres du Crest-Ensai pour leur disponibilité, leur sympathie et les échanges que nous avons eus tout au long de ma thèse. Je pense notamment à Myriam et à Samuel.

Je remercie tout particulièrement les membres du laboratoire de statistique d'enquête qui constituent une équipe soudée et dynamique où il est agréable de travailler. Merci à Cyril, Daniel, Guillaume et Mohammed.

Enfin, je remercie Guillaume Chauvet et David Haziza pour l'aide essentielle qu'ils m'ont apportée dans la rédaction de cette thèse.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Un institut public de sondages comme l'Insee a pour mission d'éclairer le débat économique et social. Pour ce faire, il collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises. La population française est vue à cet égard comme une population finie fixée. Les informations la concernant sont pour l'essentiel des statistiques descriptives de la population française ou de l'économie française estimées à partir de données collectées par enquête auprès d'échantillons probabilistes. Ainsi, le taux de chômage ou le nombre total de chômeurs sont estimés à partir de l'enquête emploi. En théorie des sondages, les statistiques descriptives sont appelées des **paramètres de population finie**. Elles sont estimées à partir des valeurs observées pour les unités échantillonnées pondérées par un jeu de poids choisi de façon à obtenir un estimateur convergent et le plus efficace possible. On appelle variables d'intérêt les caractéristiques d'intérêt mesurées par l'enquête, comme la situation par rapport à l'emploi. Pour une variable d'intérêt y, on note y_k la valeur de cette caractéristique pour l'unité k de la population finie notée $\mathcal{U}$.
En plus des données collectées par sondages, les instituts publics ont accès à des fichiers de données exhaustives ou des statistiques descriptives de la population totale qui proviennent de recensements et/ou de fichiers administratifs. Ces deux types d'information, ne provenant pas de la collecte, sont appelées **information auxiliaire**. En sondages, la mobilisation de l'information auxiliaire est d'une grande importance. A l'étape de l'échantillonnage, i.e. de la sélection de l'échantillon, l'information auxiliaire est utilisée pour construire le plan de sondage le plus optimal possible. A l'étape de l'estimation, l'information auxiliaire sert aux trois étapes de la construction des poids. D'abord, dans l'étape du calcul des poids pour obtenir un estimateur conditionnellement sans biais, ensuite dans le calcul du facteur d'ajustement pour corriger l'erreur due à la non-réponse et enfin dans le calcul du facteur d'ajustement destiné à obtenir des estimations cohérentes des totaux des variables auxiliaires.
L'inférence traditionnelle en sondage repose sur le plan de sondage, i.e. sur la loi de probabilités $p(s)$ régissant la sélection de l'échantillon s. L'information auxiliaire est utilisée à l'étape de l'échantillonnage avec l'objectif d'obtenir des estimateurs sans biais de variance minimum. Bien que la plupart des statisticiens d'enquête sont réticents à parler de modèles sur les variables d'intérêt, c'est bien l'idée d'une relation entre les variables d'intérêt et les variables auxiliaires qui guide le choix du

plan de sondage optimal (Chambers et Clark, 2012). On peut citer l'exemple des plans de sondages équilibrés qui sélectionnent des échantillons s dont les estimations des totaux des variables auxiliaires sont exactement égales aux vrais totaux et qui sont motivés par l'existence de modèles linéaires entre les variables d'intérêt et les variables auxiliaires.
L'information auxiliaire complète, en provenance d'une source administrative, peut être utilisée pour réaliser une inférence conditionnelle. Si l'estimateur du total de la variable d'intérêt sur la population $\mathcal{U}$, noté t_y, est lié à l'estimateur du total de la variable auxiliaire, t_x, alors ce premier présentera un biais dans le cadre de l'inférence conditionnelle. L'estimateur linéaire pondéré par les inverses des probabilités d'inclusion conditionnelles sera quant à lui conditionnellement sans biais.
L'information auxiliaire est également mobilisée à l'étape de l'ajustement des poids pour tenir compte du mécanisme de non-réponse. Les variables explicatives de la non-réponse sont prises, traditionnellement, parmi les variables auxiliaires. Les probabilités de réponse $\hat{p}_k$, $k \in s$, sont estimées à partir du modèle de non-réponse et permettent d'ajuster le poids initial d'une unité k par un facteur d'ajustement qui est égal à l'inverse de l'estimation de la probabilité de réponse de l'individu k.
Enfin, à l'étape du calage, toujours dans l'approche « classique » des sondages, l'information auxiliaire est utilisée pour ajuster les poids de sondages afin d'obtenir des estimations exactes des totaux des variables auxiliaires.

Dans cette thèse, on aborde des questions d'échantillonnage, d'estimation, d'inférence et de correction de la non-réponse. Pour chacune de ces problématiques, l'accent est mis sur l'utilisation optimale de l'information auxiliaire dans le but d'obtenir le meilleur estimateur linéaire pondéré possible.

Dans le chapitre 2, nous donnerons une présentation des notions de la théorie des sondages qui seront utilisées dans les chapitres suivants. Nous définirons les concepts de population finie, de modèle de superpopulation, d'estimateur Horvitz-Thompson, d'asymptotique en population finie, de paramètres complexes, de linéarisation de paramètres complexes, d'estimateurs assistés par un modèle, d'estimateurs par calage, d'estimateurs basés sur un modèle de prédiction, de non-réponse et d'inférence conditionnelle.

L'objet du chapitre 3 est d'étendre la famille des estimateurs par calage (Deville et Särndal, 1992). Actuellement la méthode du calage permet de caler sur des totaux. L'idée est de pouvoir prendre en compte dans les contraintes de calage des paramètres de population finie complexes, comme un ratio, une médiane ou une moyenne géométrique. Ce qui motive cette démarche est l'éventualité de disposer d'une information auxiliaire sous la forme d'un paramètre complexe et non sous la forme de totaux. Par exemple, le ratio sur la population totale est connu mais le total du numérateur et le total du dénominateur ne le sont pas. D'autres auteurs ont déjà abordé la question des paramètres complexes dans le cadre du calage. Särndal (2007) passe en revue un certain nombre d'entre eux. Il cite notamment les travaux de Harms et Duchesne (2006) sur l'estimation de quantiles par calage, et ceux de Plikusas (2005) et Krapavickaite et Plikusas (2006) sur les estimateurs par calage de

certaines fonctions de totaux. L'originalité de notre travail est de ramener le calage sur un paramètre complexe à un calage sur le total d'une nouvelle variable auxiliaire ad hoc. L'avantage de cette approche est de permettre l'utilisation des outils actuels de calage et de ne pas avoir à résoudre un programme d'optimisation complexe. Nous rappellerons le fonctionnement de la méthode de calage, définirons le calage sur paramètres complexes et présenterons des cas simples où le calage sur un paramètre complexe peut se ramener à un calage sur un total. Lorsque les paramètres peuvent se définir comme solution d'une équation estimante (Godambe et Thompson, 1986), nous introduirons la notion de calage sur paramètre complexe défini par une équation estimante et montrerons que l'équation de calage résultante peut être remplacée par une équation de calage sur un total.
Ce travail a été publié dans la revue Techniques d'Enquête (Lesage, 2011).

Le chapitre 4 présente un travail réalisé en collaboration avec David Haziza. Nous nous sommes intéressés à l'utilisation du calage pour corriger l'erreur liée à la non-réponse. La pratique la plus couramment utilisée pour corriger des erreurs d'échantillonnage et de non-réponse consiste à estimer dans un premier temps la probabilité de répondre des unités k à l'aide d'un modèle de non-réponse puis, dans un deuxième temps, de procéder à un calage avec les poids initiaux multipliés par un facteur d'ajustement de la non-réponse qui est l'inverse de l'estimation de la probabilité de réponse. La première étape vise à réduire le biais de non-réponse. Cet objectif est atteint à la condition de disposer du bon modèle de non-réponse et des variables explicatives de la non-réponse parmi les variables auxiliaires. La seconde étape est de nature plus cosmétique, elle permet d'assurer la cohérence entre les estimations et les vraies valeurs des totaux connus. Si la variable d'intérêt et les variables de calage sont liées, de façon linéaire, on peut espérer un gain en précision.
Une approche alternative (voir Särndal et Lundström (2005)) propose de corriger des erreurs d'échantillonnage et de non-réponse par un calage unique : c'est l'approche en une étape. Deville and Särndal (1992) ont montré qu'en absence d'erreur de non-réponse les estimateurs par calage sont asymptotiquement équivalents quelle que soit la fonction de calage utilisée et qu'ils sont sans biais. Nous montrons qu'il n'en va plus de même en présence de non-réponse et que le choix de la fonction de calage peut mener à des estimateurs dont les propriétés en termes de biais et de variance sont très différentes. Alors que le choix des variables de calage a été largement discuté dans la littérature scientifique (par exemple Särndal et Lundström (2005) et Särndal (2011)) la question du choix de la fonction de calage et les conséquences d'un mauvais choix n'ont pas été complètement explorées (à l'exception de Kott (2006) et Kott et Liao (2012)).
Dans ce chapitre, nous montrons qu'en dépit du fait que le calage n'utilise pas explicitement les probabilités de réponse, il est nécessaire d'écrire le modèle de réponse afin de choisir correctement la fonction de calage. A défaut, on s'expose à des estimateurs biaisés dont le biais peut dépasser le biais de l'estimateur non-ajusté.
Un article portant sur ce travail a été soumis (Haziza et Lesage, 2013).

Au chapitre 5, il est question d'un travail fait en commun avec David Haziza sur le calage généralisé. Cet estimateur a connu un grand intérêt ces dernières années compte

tenu de ses capacités supposées pour corriger le biais dû à la non-réponse (voir Deville (2002), Sautory (2003), Särndal et Lundström (2005), Kott (2006, 2009), Chang et Kott (2008), Kott et Chang (2010) et Kott et Liao (2012)). Le calage généralisé est une version plus sophistiquée du calage où la fonction de calage dépend d'un vecteur de variables instrumentales qui peuvent être différentes des variables de calage et n'être connues que sur l'échantillon des répondants. Les variables instrumentales ont vocation à être des variables explicatives de la non-réponse. Ainsi, l'avantage du calage généralisé sur le calage simple pour corriger la non-réponse réside dans la possibilité de corriger des mécanismes de non-réponse classés non-MAR (non Missing at Random) conditionnellement aux variables auxiliaires. Ces cas incluent la situation où la variable d'étude est elle-même explicative de la non-réponse. On met en évidence, selon le principe de double robustesse assez classique en sondage, que l'estimateur par calage généralisé est sans biais si le facteur d'ajustement correspondant à la fonction de calage estime l'inverse des probabilités de réponse de façon convergente ou si la variable d'intérêt et les variables de calage suivent des modèles de régression linéaire en fonction des variables instrumentales. La seconde situation, appelée approche modèle, nécessite également que les variables de calage n'interviennent pas dans le modèle de non-réponse en plus des variables instrumentales. Dans ces conditions, il reste à être prudent sur le choix des variables de calage, car si celles-ci sont trop faiblement corrélées aux variables instrumentales, la variance est amplifiée (Osier, 2012).
Par ailleurs, on montre que si les variables de calage interviennent dans le modèle de non-réponse en plus des variables instrumentales, alors l'estimateur par calage généralisé est biaisé. En outre, comme pour la variance, ce biais est amplifié pour des variables de calage faiblement corrélées aux variables instrumentales.
Nous concluons que l'utilisation du calage généralisée pour traiter la non-réponse est extrêmement délicate, car il est impossible de s'appuyer sur une modélisation de la non-réponse pour choisir les bonnes variables explicatives de la non-réponse et la forme fonctionnelle du modèle de non-réponse.
Un article portant sur ce travail a été soumis (Lesage et Haziza, 2013).

Le chapitre 6 présente un travail sur l'inférence conditionnelle qui s'inscrit dans la recherche d'un estimateur linéaire pondéré sans biais et efficace par rapport au plan de sondage. Les estimateurs utilisés en sondage ont souvent fait l'objet de critiques de la part d'auteurs préférant une approche basée sur un modèle statistique décrivant les variables d'intérêt (voir par exemple Royall et Cumberland (1981)). Ces auteurs ont montré que connaissant la relation entre la variable d'intérêt et les variables auxiliaires, on pouvait détecter au vu de l'échantillon tiré, si l'estimateur proposé par la théorie des sondages était biaisé. Ils affirmaient, non sans raison, qu'il était aberrant d'invoquer le caractère non-biaisé par rapport au plan de sondage dans un cas où il était manifeste que l'unique échantillon dont on disposait offrait une image déformée de la population étudiée.
L'inférence conditionnelle en sondage (Rao (1985) et Tillé (1998 et 1999)) et la proposition d'estimateur conditionnellement sans biais ont permis de lever ces critiques tout en conservant une inférence par rapport au plan de sondage. Toutefois, les calculs exacts des lois conditionnelles ou même des probabilités d'inclusion condi-

tionnelles deviennent rapidement complexes lorsque le plan de sondage n'est pas à probabilités égales. Nous avons montré qu'il était possible de calculer les probabilités d'inclusion conditionnelles dans le cas d'un sondage à probabilités inégales de taille fixe : le sondage de Poison conditionnel à la taille de l'échantillon. On a également proposé une nouvelle méthode d'estimation des probabilités d'inclusion conditionnelles à partir de méthodes de simulations Monte Carlo. Auparavant, il était proposé d'utiliser des hypothèses de normalité asymptotique pour construire des estimateurs conditionnellement sans biais qui étaient équivalents à l'estimateur optimal de Montanari (Rao, 1985). Avec la méthode d'estimation par Monte Carlo, on a l'avantage de pouvoir traiter les problèmes d'unités influentes résultant de valeurs extrêmes (outliers) et de « saut de strate ». Nos simulations montrent que les probabilités d'inclusion des unités influentes peuvent être ramenée à 1.
Ce travail est en cours de révision (Coquet et Lesage, 2012).

Au chapitre 7, il est question d'un travail fait en commun avec Guillaume Chauvet et David Haziza sur l'échantillonnage équilibré. L'échantillonnage équilibré est à l'échantillonnage ce que le calage est à l'estimation. Un plan de sondage équilibré offre des estimateurs par expansion calés sur les totaux des variables auxiliaires qui ont servi à l'équilibrage. Il existe de nombreux algorithmes de tirage d'échantillons qui fournissent des plans de sondage équilibrés ou approximativement équilibrés. Nous nous intéressons plus particulièrement à deux algorithmes : la méthode du Cube (Deville et Tillé, 2004) et le tirage réjectif de Fuller (2009).
Dans le cas du tirage réjectif, il s'agit simplement d'utiliser un plan initial qui n'est pas équilibré (un plan de sondage aléatoire simple sans remise par example) et d'itérer son algorithme jusqu'à obtention d'un échantillon qui remplisse la condition d'équilibrage. Fuller (2009), ainsi que Legg et Yu (2010), ont montré que l'estimateur par la régression associé au plan rejectif était convergent.
Nous montrons que l'inconvénient de cette méthode par rapport à la méthode du cube est que les probabilités d'inclusion ne sont plus forcement égales aux probabilités de sélection initiales pour des tailles d'échantillon finies et qu'en conséquence l'estimateur par la régression peut être biaisé lorsque la variable d'intérêt ne suit pas un modèle de régression linéaire en fonction des variables auxiliaires.
Nous proposons d'utiliser un estimateur par expansion avec des probabilités d'inclusion estimées afin d'obtenir un estimateur sans biais. Nous comparons deux méthodes d'estimation des probabilités d'inclusion : une méthode de simulations Monte Carlo (Fattorini(2006) et Thompson et Wu (2008)) et une méthode reposant sur des expansions d'Edgeworth (Hájek et Dupač, 1981).

On trouve en annexe de la thèse un travail sur le calage sur les paramètres provenant d'une analyse en composantes principales (ACP). L'objectif visé était de pondérer l'échantillon de façon a retrouver les mêmes composantes principales et les mêmes inerties que celles obtenues à partir de l'ACP sur la population totale $\mathcal{U}$.
A l'origine, ce travail se voulait une application du calage sur paramètre complexe présenté au chapitre 2. Finalement, il est apparu que cette méthode relevait davantage des questions de multi-colinéarité et de parcimonie que de la question du calage sur un paramètres complexes.

Bibliographie

Chambers, R. and Clark, R. (2012). *An introduction to model-based survey sampling with applications*. OUP Oxford.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95(3) :555–571.

Coquet, F. and Lesage, E. (2012). Conditional inference with a complex sampling : exact computations and monte carlo estimations. *In revision*.

Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des journées de méthodologie statistique*, pages 4–20.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418) :376–382.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling : the cube method. *Biometrika*, 91(4) :893–912.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs : A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2) :269–278.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4) :933–944.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population : Their relationship and estimation. *International Statistical Review*, 54 :127–138.

Hájek, J. and Dupač, V. (1981). *Sampling from a finite population*, volume 37. M. Dekker.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32 :37–52.

Haziza, D. and Lesage, E. (2013). A discussion of weighting procedures in the presence of unit nonresponse. *Submitted for publication*.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2) :133.

Kott, P. S. (2009). Calibration weighting : Combining probability samples and linear prediction models. *Handbook of Statistics, Sample Surveys : Inference and Analysis*, 29B :55–82.

Kott, P. S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. In *Survey Research Methods*, volume 6, pages 105–111.

Krapavickaite, D. and Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica*, 16 :347–364.

Legg, J. C. and Yu, C. L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, pages 69–79.

Lesage, E. (2011). The use of estimating equations to perform a calibration on complex parameters. *Survey Methodology*, 37(1) :103–108.

Lesage, E. and Haziza, D. (2013). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. *Submitted for publication*.

Osier, G. Dealing with non-ignorable non-response using generalised calibration : A simulation study based on the luxemburgish household budget survey. *Economie et Statistiques, Working papers du STATEC*, (65).

Plikusas, A. (2006). Non-linear calibration. In *Proceedings, Workshop on survey sampling*, Venspils, Latvia. Riga : Central Statistical Bureau of Latvia.

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11 :15–31.

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76(373) :66–77.

Sautory, O. (2003). Calmar 2 : une nouvelle version du programme calmar de redressement d'échantillon par calage. In *Recueil : Symposium de Statistique Canada*.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2) :113–135.

Särndal, C.-E. (2011). Three factors to signal non-response bias with applications to categorical auxiliary variables. *International Statistical Review*, 79(2) :233–254.

Särndal, C.-E., Lundström, S., and Wiley, J. (2005). *Estimation in surveys with nonresponse*. Wiley Hoboken, NJ.

Thompson Mary, E. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34(1) :3–10.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities : Simple random sampling. *International Statistical Review*, 66 :303–322.

Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities : comlex design. *Survey Methodology*, 25(1) :57–66.

Chapitre 2

Inférence en population finie

2.1 Contexte d'un sondage probabiliste

2.1.1 La population d'étude et la statistique d'intérêt

On considère une population finie de taille N dont les unités sont identifiées par un indice k appartenant à l'ensemble $\mathcal{U} = \{1, 2, ..., N\}$.
On s'intéresse principalement à des statistiques descriptives de la population finie construites à partir d'un vecteur de q caractéristiques d'intérêt (ou variables d'intérêt) noté $\boldsymbol{y} = (y_1, y_2, \ldots, y_q)^\top$. Ces statistiques descriptives, encore appelées paramètres de la population finie, sont par exemple le total d'une caractéristique d'intérêt $t_{y1} = \sum_{k\in\mathcal{U}} y_{k,1}$, sa moyenne $\mu_{y1} = t_{y1}/N$ ou encore un ratio des totaux de deux variables d'intérêt $R_{1,2} = \dfrac{t_{y2}}{t_{y1}}$.

Nous n'aborderons pas ici l'estimation de paramètres de modèles statistiques associés à la population $\mathcal{U}$ (modèles économétriques par exemple). Ces paramètres sont qualifiés de paramètres de superpopulation par opposition aux paramètres de population finie.

2.1.2 Le modèle de superpopulation

En théorie des sondages, même si on s'intéresse davantage aux observations qu'aux modèles statistiques qui les engendrent, il est tout de même utile, voire nécessaire, d'avoir recours à des modèle statistiques en présence d'erreurs non dues à l'échantillonnage (erreurs de non-réponse, erreurs de couverture et erreurs de mesure).
On note $\boldsymbol{x}_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,p})^\top$ le vecteur des valeurs pour l'unité k des p caractéristiques $(x_1, x_2, \ldots, x_p)$ qui ne font pas partie des caractéristiques d'intérêt et $\boldsymbol{y}_k = (y_{k,1}, y_{k,2}, \ldots, y_{k,q})^\top$ les valeurs des q variables d'intérêt de l'unité k.

Dans le cadre du modèle, appelé **modèle de superpopulation**, les vecteurs des observations $\left(\boldsymbol{x}_k^\top,\ \boldsymbol{y}_k^\top\right)^\top$, $k \in \mathcal{U}$, sont les réalisations des vecteurs aléatoires i.i.d. $\left(\boldsymbol{X}_k^\top,\ \boldsymbol{Y}_k^\top\right)^\top$, où $\boldsymbol{X}_k = (X_{k,1}, X_{k,2}, \ldots, X_{k,p})^\top$ et $\boldsymbol{Y}_k = (Y_{k,1}, X_{k,2}, \ldots, Y_{k,q})^\top$. On

note $\mathcal{F}_N = \left\{ \left(\boldsymbol{X}_k^\top,\ \boldsymbol{Y}_k^\top \right) = \left(\boldsymbol{x}_k^\top,\ \boldsymbol{y}_k^\top \right), k \in \mathcal{U} \right\}$ l'événement correspondant à la réalisation de notre population finie.

L'espérance et la variance correspondant au modèle de superpopulation sont notées $\mathbb{E}_m(\cdot)$ et $\mathbb{V}_m(\cdot)$ et qualifiées d'espérance et de variance par rapport au modèle (de superpopulation).

2.1.3 Le plan de sondage

2.1.3.1 Le plan de sondage

Pour calculer les paramètres d'une population finie, il est courant, pour des raisons économiques, de procéder à une enquête par sondage plutôt qu'à un recensement. Ainsi, les valeurs d'une variable d'intérêt y sont collectées uniquement pour un sous-ensemble de la population finie $\mathcal{U}$. Ce sous-ensemble de taille n est appelé **échantillon** et est noté s. D'un point de vue pratique, les unités échantillonnés sont choisies dans une liste, appelée **base de sondage**, qui contient au minimum les coordonnées de chaque unité k de $\mathcal{U}$.

L'échantillon s est sélectionné aléatoirement suivant une loi de probabilité $P(s)$ appelée **plan de sondage**. Le plan de sondage est construit en prenant en compte les caractéristiques des unités de la population finie. Si on utilise le modèle de superpopulation, le plan de sondage correspond à une loi de probabilité conditionnelle à la population finie (autrement dit à la réalisation de l'événement $\mathcal{F}_N$).

On note $\mathcal{S} = \{s \subset \mathcal{U},\ P(s) > 0\}$ l'ensemble des échantillons possibles.

L'espérance et la variance d'une variable aléatoire $\hat{\theta}$ par rapport au plan de sondage sont définies par :

$$\mathbb{E}_P(\hat{\theta}) = \sum_{s \in \mathcal{S}} P(s)\hat{\theta}(s) \tag{2.1}$$

et

$$\mathbb{V}_P(\hat{\theta}) = \mathbb{E}_P(\hat{\theta}^2) - \left\{ \mathbb{E}_P(\hat{\theta}) \right\}^2. \tag{2.2}$$

On définit I_k les variables aléatoires indicatrices d'appartenance à l'échantillon pour toutes les unités $k \in \mathcal{U}$:

$$I_k = \begin{cases} 1 & \text{si } k \in s, \\ 0 & \text{si } k \notin s. \end{cases}$$

On définit également, pour chaque unité k, sa probabilité d'inclusion dans l'échantillon à l'ordre 1 notée π_k :

$$\begin{aligned} \pi_k &= P\left([I_k = 1]\right) \\ &= \mathbb{E}_P\left(I_k\right). \end{aligned}$$

La probabilité d'inclusion à l'ordre 2 (ou jointe) de deux unités k et l est définie par :

$$\begin{aligned} \pi_{k,l} &= P\left([I_k = 1] \cap [I_l = 1]\right) \\ &= \mathbb{E}_P\left(I_k I_l\right). \end{aligned}$$

2.1.3.2 Exemples de plans de sondage simples

Le plan de sondage de taille fixe le plus simple est le plan de sondage aléatoire simple sans remise qui permet de tirer des échantillons de taille fixe n de façon équiprobable :

$$P(s) = \binom{N}{n}^{-1} \mathbb{1}_{(|s|=n)}.$$

Les probabilités d'inclusion valent $\pi_k = n/N$ et les probabilités d'inclusion jointes $\pi_{k,l} = \dfrac{n(n-1)}{N(N-1)}$ pour les couples $(k,\ l)$ tels que $k \neq l$.

Parmi les plans de sondage simples de taille aléatoire, on peut citer le sondage de Poisson où les unités sont sélectionnées indépendamment suivant des lois de Bernoulli de paramètres π_k :

$$P(s) = \prod_{k \in \mathcal{U}} \pi_k^{I_k} (1 - \pi_k)^{1 - I_k}.$$

Les probabilités d'inclusion jointes valent $\pi_{k,l} = \pi_k\ \pi_l$ pour $k \neq l$.

2.1.4 L'estimateur Horvitz-Thompson d'un total

En sondages, on appelle **stratégie** le couple formé par le plan de sondage et l'estimateur du paramètre d'intérêt. L'objectif du statisticien d'enquête est de choisir la stratégie qui fournira l'estimation du paramètre la plus précise, i.e. l'intervalle de confiance du paramètre le plus étroit.
On s'intéresse à l'estimation du total de la variable y sur la population $\mathcal{U}$

$$t_y = \sum_{k \in \mathcal{U}} y_k.$$

L'estimateur de Horvitz-Thompson du total t_y (Hovitz et Thompson, 1952) est défini par

$$\hat{t}_{y,\pi} = \sum_{k \in \mathcal{U}} I_k \frac{y_k}{\pi_k} = \sum_{k \in \mathcal{U}} d_k I_k y_k, \tag{2.3}$$

où $d_k = \pi_k^{-1}$ est le **poids d'échantillonnage** de l'unité k, ou encore poids de sondage. Cet estimateur est également appelé le π-estimateur ou l'estimateur par expansion.

Si $\pi_k > 0$ pour tout $k \in \mathcal{U}$, alors l'estimateur (2.3) est sans biais par rapport au plan :

$$\mathbb{E}_P\left(\hat{t}_{y,\pi}\right) = t_y.$$

Sa variance (par rapport au plan) vaut

$$\begin{aligned}\mathbb{V}_P\left(\hat{t}_{y,\pi}\right) &= \sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}\left\{\mathbb{E}_P(I_kI_l)-\mathbb{E}_P(I_k)\mathbb{E}_P(I_l)\right\}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\\ &= \sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}\left(\pi_{k,l}-\pi_k\pi_l\right)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \qquad (2.4)\end{aligned}$$

Un estimateur de la variance (2.4) est donné par

$$\hat{\mathbb{V}}\left(\hat{t}_{y,\pi}\right)=\sum_{k\in s}\sum_{l\in s}\frac{\left(\pi_{k,l}-\pi_k\pi_l\right)}{\pi_{k,l}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \qquad (2.5)$$

Si $\pi_{k,l}>0$ pour tous les couples $(k,l)\in\mathcal{U}^2$, alors l'estimateur de variance (2.5) est sans biais

$$\mathbb{E}_P\left\{\hat{\mathbb{V}}\left(\hat{t}_{y,\pi}\right)\right\}=\mathbb{V}_P\left(\hat{t}_{y,\pi}\right).$$

Pour un plan de sondage aléatoire simple sans remise, (2.4) se simplifie pour donner

$$\mathbb{V}_P\left(\hat{t}_{y,\pi}\right) = N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}, \qquad (2.6)$$

où $S_y^2=\dfrac{1}{N-1}\displaystyle\sum_{k\in\mathcal{U}}(y_k-\mu_y)^2$ est la dispersion de la variable y sur $\mathcal{U}$
et (2.5) devient :

$$\hat{\mathbb{V}}_P\left(\hat{t}_{y,\pi}\right) = N^2(1-\frac{n}{N})\frac{s_y^2}{n}, \qquad (2.7)$$

où $s_y^2=\dfrac{1}{n-1}\displaystyle\sum_{k\in s}(y_k-\mu_y)^2$ est la dispersion de la variable y sur s.

La grande popularité de cet estimateur est due à sa simplicité et au fait que ce soit un **estimateur linéaire pondéré**. Cet estimateur se résume donc à la connaissance d'un jeu de poids unique pour les unités échantillonnées qui servent à l'estimation de n'importe quelle variable d'intérêt de l'enquête. Chaque individu interrogé a un poids supérieur à 1, c'est-à-dire que l'individu compte pour lui-même plus d_k-1 autres individus non sélectionnés.

La propriété fondamentale de l'estimateur par expansion est qu'il est sans biais sous le plan de sondage quel que soit $n\geq 1$. Cette propriété repose sur le fait que pour tout unité k de $\mathcal{U}$, $\mathbb{E}_P\left(d_kI_k\right)=1$, c'est-à-dire que les poids valent tous 1 en moyenne par rapport au plan.

On verra dans le chapitre 6 qu'on peut construire, à partir d'une approche conditionnelle, d'autres jeux de poids w_k qui vérifient la propriété :

$$\forall k,\quad \mathbb{E}_P\left(w_kI_k\right)=1.$$

2.1.4.1 L'inférence en sondages

L'inférence en sondage consiste à calculer une estimation ponctuelle pour le total t_y, une estimation de variance de l'estimateur utilisé et enfin à donner un intervalle de confiance à 95% pour le paramètre t_y. Pour construire cet intervalle de confiance, on a recours à des arguments asymptotiques qui vont être présentés dans la Section 2.1.5.

2.1.4.2 Information auxiliaire

Les instituts de statistique publics disposent d'autres sources d'information sur la population $\mathcal{U}$ que l'enquête menée spécifiquement pour collecter les variables d'intérêt. Des recensements et des fichiers administratifs fournissent des informations extérieures à l'enquête ; c'est ce qu'on appelle **l'information auxiliaire**. Cet information auxiliaire peut porter sur la connaissance de statistiques descriptives globales portant sur la population $\mathcal{U}$ où sur la connaissance de caractéristiques pour toutes les unités de la population. Dans les deux cas de figure, on appelle variables auxiliaires les p caractéristiques pour lesquelles on a de l'information auxiliaire et on note $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ le vecteur de ces variables. Lorsque l'information auxiliaire est disponible avant la collecte, elle peut servir à construire le plan de sondage. Elle peut également être utilisée à l'étape de l'estimation dans des procédures d'ajustement des poids de sondage.

2.1.4.3 Exemples de plans de sondages complexes

Dans le domaine de la statistique publique, les plans de sondage sont en général plus complexes que ceux évoqués à la Section 2.1.3.2. Dans les enquêtes sociales et démographiques, ce sont souvent des plans de sondage à plusieurs degrés : on sélectionne par exemple au premier degré un échantillon de communes puis pour chaque commune sélectionnée, on sélectionne, au deuxième degré, un échantillon de ménages. Les raisons de cette complexité sont en partie d'ordre économique : on tire d'abord des communes puis des ménages afin de concentrer géographiquement les échantillons dans le but de diminuer les frais de déplacement des enquêteurs. Mais il existe également des raisons pratiques : on ne dispose pas toujours d'une base de sondage qui donne directement accès aux éléments de $\mathcal{U}$.
On emploie également souvent la technique de **stratification** lorsque la population $\mathcal{U}$ n'est pas homogène pour la variable d'intérêt y mais qu'elle peut être découpée en sous-populations qui elles le sont. Ces sous-populations homogènes sont appelées des strates et le plan de sondage stratifié consiste à utiliser des plans de sondage indépendants dans ces strates. Une des propriétés intéressantes de la stratification est qu'avec des plans de sondage aléatoires simples dans chacune des strates, l'estimateur par expansion donne des estimateurs des tailles des strates égales à leurs vraies valeurs ; on a donc un plan de sondage équilibré sur les tailles des strates. Cette propriété conduit à un gain en précision par rapport à un plan de sondage aléatoire simple (sans remise) pour des variables d'intérêt qui suivent un modèle d'analyse de la variance en fonction de la variable de stratification.

Une autre méthode souvent utilisée est **l'échantillonnage équilibré**. Avec un plan de sondage équilibré, l'estimateur par expansion et la vraie valeur du vecteur des totaux $\boldsymbol{t}_x$ des variables d'équilibrage $\boldsymbol{x}$ sont égaux :

$$\hat{\boldsymbol{t}}_{x,\pi} = \boldsymbol{t}_x, \tag{2.8}$$

où $\hat{\boldsymbol{t}}_{x,\pi} = \sum_{k\in\mathcal{U}} I_k \frac{\boldsymbol{x}_k}{\pi_k}$.

Parmi les algorithmes de tirage qui permettent d'obtenir un plan de sondage équilibré, on peut citer la méthode du cube (Deville et Tillé, 2004). En utilisant un plan de sondage équilibré, les instituts de sondage publics cherchent d'abord à obtenir des estimateurs des totaux connus cohérents avec les vrais totaux, mais également des estimations plus précises lorsque la variable d'intérêt suit un modèle linéaire en fonction des variables d'équilibrage.

2.1.5 Cadre asymptotique

En théorie des sondages, dans le cadre d'une inférence par rapport au plan (qu'on appelle également approche par randomisation) les estimateurs ont des lois de probabilité compliquées difficile à expliciter (même pour un estimateur simple comme l'estimateur Horvitz-Thompson). On ne peut donc pas faire de tests ou calculer des intervalles de confiance à distance finie (i.e. pour des tailles d'échantillon finies) et il faut recourir à une approche asymptotique pour obtenir une loi approchée de l'estimateur $\hat{t}_{y,\pi}$ et ainsi construire un intervalle de confiance. Pour des estimateurs complexes comme les estimateurs assistés par un modèle (voir la Section 2.4), les arguments asymptotiques sont également nécessaires pour proposer des estimateurs du biais et de la variance ; on reviendra sur ce sujet à la Section 3.4 qui traite de la méthode de linéarisation des estimateurs.

On trouve de nombreux théorèmes concernant l'asymptotique en sondage car les hypothèses et les démonstrations varient en fonction des plans de sondage, des estimateurs et de la modélisation de la séquence des populations finies dont les effectifs tendent vers l'infini ; voir par exemple Isaki et Fuller (1982), et Fuller (2009a, P.35-56).

On construit un cadre de modélisation dans lequel on peut faire tendre la taille de la population N et l'espérance de la taille d'échantillon n_N vers l'infini. On considère une suite de populations finies indexées par N et de taille N (par commodité) : $\mathcal{U}_N = (1,\ \ldots,\ N)$.
On note $\mathcal{F}_N = (y_{1,N},\ \ldots,\ y_{N,N})$ la N^{ieme} population finie.
La séquence $(\mathcal{F}_N;\ N \in \mathbb{N})$ peut être modélisée de deux façons différentes. Soit l'ensemble $\mathcal{F}_N$ est un vecteur de valeurs fixées dans une séquence elle-même fixée. Soient les valeurs $y_{k,N}$ sont des réalisations de variables aléatoires $Y_{k,N}$.

Exemple 1. *Prenons l'exemple (Fuller, 2009a) d'un sondage aléatoire simple sans remise de taille* $n_N = [fN]$ *où* $0 < f < 1$. *Ajoutons la condition*

$$\lim_{N\to\infty} N^{-1} \sum_{k=1}^{N} \left(y_{k,N}, y_{k,N}^2\right) = (\theta_1, \theta_2) \qquad ,$$

où (θ_1, θ_2) *sont des constantes telles que* $\theta_2 - \theta_1^2 > 0$.

Alors on a

$$N^{-1}\left(\hat{t}_{y,\pi} - t_{y,N}\right) \mid \mathcal{F}_N = O_P\left(\frac{1}{n_N^{1/2}}\right). \tag{2.9}$$

On définit le degré α *d'un paramètre* θ_N *comme la valeur maximale* $\alpha \in \mathbb{Q}$ *telle que* $N^{-\alpha}\theta_N$ *soit borné à partir d'un certain rang de* N. *Le degré de* t_y *est* $\alpha = 1$.
En théorie des sondages, on dit qu'un estimateur $\hat{\theta}_N$ *d'un paramètre* θ_N *de degré* α *est* $\sqrt{n}$*-convergent si et seulement si* $\hat{\theta} - \theta = O_P\left(N^{\alpha} n_N^{-1/2}\right)$.
Le résultat (2.9) implique que l'estimateur $\hat{t}_{y,\pi}$ *est* $\sqrt{n}$*-convergent sous le plan (on dira également* $\sqrt{n}$*-consistant).*

Pour construire un intervalle de confiance, il faut un résultat de type « théorème de la limite centrale ». Sous certaines conditions (portant notamment sur les moments d'ordre 4 de y, *voir Fuller(2009a)), on a :*

$$\frac{\hat{t}_{y,\pi} - t_{y,N}}{N\left\{\left(1 - \frac{n_N}{N}\right)\frac{s_{y,N}^2}{n_N}\right\}^{-1/2}} | \mathcal{F}_N \xrightarrow[N\to+\infty]{\mathcal{L}} N(0,1).$$

Dans la pratique, quel que soit le plan de sondage, on suppose qu'on a les conditions suffisantes pour pouvoir écrire :

$$\frac{\hat{t}_{y,\pi} - t_{y,N}}{\hat{\mathbb{V}}\left(\hat{t}_{y,\pi}\right)^{-1/2}} | \mathcal{F}_N \xrightarrow[N\to+\infty]{\mathcal{L}} N(0,1). \tag{2.10}$$

Ce dernier résultat permet de construire un intervalle de confiance pour le paramètre t_y en utilisant les quantiles de la loi normale. Ainsi, l'intervalle de confiance à 95% vaut :

$$IC_{95\%}(\hat{t}_{y,\pi}) = \left[\hat{t}_{y,\pi} - 1.96\sqrt{\hat{\mathbb{V}}\left(\hat{t}_{y,\pi}\right)},\ \hat{t}_{y,\pi} + 1.96\sqrt{\hat{\mathbb{V}}\left(\hat{t}_{y,\pi}\right)}\right] \tag{2.11}$$

Remarque 1. *Tous ces résultats sont obtenus pour des distributions de* y *avec des moments empiriques finis et des probabilités d'inclusion qui vérifient pour tout* $k \in \mathcal{U}$ *:* $0 < \alpha_{inf} \leq N(n\pi_k)^{-1} \leq \alpha_{sup}$, *où* α_{inf} *et* α_{sup} *sont deux nombres réels. Ces hypothèses peuvent être mises à mal en présence de valeurs extrêmes ou d'unités influentes.*

2.2 Estimation de paramètres complexes

Les paramètres de population finie sont parfois plus complexes que des totaux. Par exemple, on peut s'intéresser à des ratios ou à des médianes. Ces paramètres complexes peuvent être définis explicitement et/ou implicitement comme des solutions d'équations estimantes. Dans les deux cas, il faut être en mesure de définir un estimateur du paramètre d'intérêt, un estimateur de son biais et un estimateur de sa variance.

2.2.1 Estimation de paramètres définis explicitement par une fonction de totaux

Soit un paramètre d'intérêt θ qui se présente sous la forme d'une fonction de q totaux $t_{y1},\ \ldots,\ t_{yq}$:

$$\theta = f\left(t_{y1}, \ldots,\ t_{yq}\right),$$

où $f(\cdot, \ldots, \cdot)$ est une fonction différentiable.
Un estimateur classique de θ est l'estimateur par substitution (également appelé **Plug-in estimateur**) :

$$\hat{\theta}_\pi = f\left(\hat{t}_{y1,\pi}, \ldots,\ \hat{t}_{yq,\pi}\right). \tag{2.12}$$

Il s'agit simplement de la fonction $f(\cdot, \ldots, \cdot)$ où les totaux $t_{y,j}$ sont remplacés par leurs estimateurs Horvitz-Thompson, $\hat{t}_{yj,\pi} = \sum_{k\in s} d_k y_{kj}$ (Särndal et al., 1992).

Pour démontrer la convergence de l'estimateur (2.12) et construire un intervalle de confiance, nous aurons recours à des méthodes de linéarisation que nous verrons à la Section 3.4.

2.2.2 Estimation de paramètres définis implicitement par des équations estimantes

Un paramètre défini implicitement est souvent un paramètre très complexe comme l'indice de Gini, qui nécessite la résolution d'un système de plusieurs équations estimantes. En effet, dans le cas de l'indice de Gini, il est nécessaire d'estimer de façon concomitante les quantiles de la variable d'intérêt. Dans la mesure où ces paramètres ne sont pas des paramètres d'intérêt, ils sont qualifiés de paramètres de nuisance (Binder, 1991).

Pour un vecteur de paramètres $\boldsymbol{\theta}$ définis implicitement par un système d'équations estimantes linéaires sur $\mathcal{U}$ (Godambe et Thompson, 1986)

$$\sum_{k\in\mathcal{U}} \boldsymbol{\Phi}(y_k, \boldsymbol{x}_k, \boldsymbol{\theta}) = \mathbf{0},$$

on associe un vecteur d'estimateurs $\hat{\boldsymbol{\theta}}_{ee}$ qui est solution d'un système d'équations estimantes linéaires sur s pondérés par les poids d'échantillonnage d_k :

$$\sum_{k\in s} d_k \boldsymbol{\Phi}\left(y_k, \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}_{ee}\right) = \mathbf{0}.$$

Prenons l'exemple simple de l'estimation de la moyenne μ_y. La moyenne μ_y peut être définie comme solution de l'équation estimante

$$\sum_{k\in U}(y_k - \mu_y) = 0.$$

Son estimateur par équation estimante $\hat{\mu}_{y,ee}$ est solution de l'équation

$$\sum_{k\in s} d_k(y_k - \hat{\mu}_{y,ee}) = 0.$$

Il peut s'écrire de façon explicite :

$$\hat{\mu}_{y,ee} = \frac{\sum_{k\in s} d_k y_k}{\sum_{k\in s} d_k},$$

aussi appelé estimateur de Hajek (1971). On peut estimer avec cette méthode des quantiles, des ratios et des paramètres de modèles linéaires ou non-linéaires. Nous y reviendrons dans le chapitre 3.

Pour démontrer la convergence de l'estimateur $\hat{\boldsymbol{\theta}}_{ee}$ et construire un intervalle de confiance, on aura recours à des méthodes de linéarisation qui seront présentées à la Section 3.4.

2.3 Linéarisation

En utilisant les propriétés asymptotiques de l'estimateur par expansion présentées à la Section 2.1.5 et des développements de Taylor, on construit, pour chaque estimateur d'un paramètre complexe, un estimateur par expansion qui lui est asymptotiquement équivalent et dont la variance est employée pour construire un intervalle de confiance.

Les techniques de linéarisation ont été introduites en théorie des sondages par Woodruff (1971) et développées par Binder (1983, 1996), Binder et Patak (1994), Wolter (1985), Deville (1999) et Demnati et Rao (2004).

2.3.1 Estimateur par substitution

Donnons les résultats principaux. L'objectif est de trouver un estimateur linéaire équivalent à l'estimateur (non-linéaire)

$$\hat{\theta} = f\left(\hat{t}_{y1}, \cdots, \hat{t}_{yq}\right)$$

du paramètre

$$\theta = f\left(t_{y1}, \cdots, t_{yq}\right),$$

où le paramètre θ est de degré α et les $\hat{t}_{yj}$ sont des estimateurs linéaires convergents de type estimateurs par expansion (mais ce pourrait être des estimateurs par calage

également).

On réécrit l'expression fonctionnelle de θ sous la forme :

$$\theta = N^{\alpha} g\left(t_{y1}/N, \cdots, t_{yq}/N\right),$$

où la fonction $g(\cdot)$ est une fonction deux fois continûment dérivable au voisinage du point $(t_{y1}/N, \cdots, t_{yq}/N)$.
On a donc

$$\hat{\theta} = N^{\alpha} g\left(\hat{t}_{y1}/N, \cdots, \hat{t}_{yq}/N\right).$$

On définit la variable linéarisée de θ par

$$\begin{aligned} v_k &= \boldsymbol{y}_k^T N^{\alpha-1} \boldsymbol{g}^{'}\left(t_{y1}/N, \cdots, t_{yp}/N\right) \\ &= \boldsymbol{y}_k^T \boldsymbol{f}^{'}\left(t_{y1}, \cdots, t_{yp}\right), \end{aligned} \tag{2.13}$$

où $\boldsymbol{f}^{'}(\cdot, \cdots, \cdot)$ (respectivement $\boldsymbol{g}^{'}(\cdot, \cdots, \cdot)$) est le vecteur des dérivées partielles de $f(\cdot)$ (respectivement $g(\cdot)$).

Sous certaines conditions Deville (1999) et Fuller (2009a), on a les résultats asymptotiques suivants :

$$N^{-\alpha}\left(\hat{\theta} - \theta\right) = N^{-\alpha}\left(\hat{t}_v - t_v\right) + O_p\left(\frac{1}{n}\right), \tag{2.14}$$

et

$$\frac{\hat{\theta} - \theta}{\hat{\mathbb{V}}\left(\hat{t}_{v,\pi}\right)^{-1/2}} | \mathcal{F}_N \xrightarrow[N \to +\infty]{\mathcal{L}} N(0,1). \tag{2.15}$$

Le théorème de la limite centrale (2.15) permet de construire un intervalle de confiance pour le paramètre θ.

Remarque 2. *On définit parfois la notion de variance de Taylor de $\hat{\theta}$ et on note*

$$A\mathbb{V}\left(\hat{\theta}\right) = \mathbb{V}_P\left(\hat{t}_v\right).$$

Cette variance de Taylor sert souvent de variance approchée de $\mathbb{V}_P(\hat{\theta})$ or il faut être prudent car les résultats présentés ici ne sont pas suffisants pour établir l'équivalence entre la variance de $\hat{\theta}$ et la variance de Taylor de $\hat{\theta}$.

2.3.2 Estimateur défini par équation implicite

Binder (1983) a montré que la technique de linéarisation permet également de définir les variables linéarisées d'un vecteur de paramètres d'intérêt définis de manière implicite par un système d'équations estimantes du type

$$\boldsymbol{T}(\boldsymbol{\theta}) = \sum_{k \in \mathcal{U}} \boldsymbol{\Phi}\left(y_k, \boldsymbol{x}_k, \boldsymbol{\theta}\right) = \boldsymbol{T_0},$$

où $\boldsymbol{\theta}$ converge (par rapport au modèle de superpopulation) vers le vecteur de paramètres $\boldsymbol{\theta}_0$ et $\boldsymbol{\Phi}$ est continument dérivable au voisinage de $\boldsymbol{\theta}_0$.

On définit l'estimateur $\hat{\boldsymbol{\theta}}_{ee}$ comme solution des équations

$$\sum_{k \in s} w_k \boldsymbol{\Phi}\left(y_k, \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}_{ee}\right) = \boldsymbol{T_0}, \tag{2.16}$$

où les w_k sont des poids ; par exemple les poids de sondage d_k mais ce peut être également des poids de calage.

Le vecteur des variables linéarisées de $\boldsymbol{\theta}$ vaut

$$\boldsymbol{v}_k = -\left[\boldsymbol{H}\left(\boldsymbol{\theta}\right)\right]^{-1} \boldsymbol{\Phi}\left(y_k, \boldsymbol{x}_k, \boldsymbol{\theta}\right),$$

où $\boldsymbol{H}\left(\boldsymbol{\theta}\right) = \partial \boldsymbol{T}(\theta)/\partial \boldsymbol{\theta}^\top$.

On obtient des propriétés asymptotiques identiques à celles de la section précédente pour chaque paramètre du vecteur $\hat{\boldsymbol{\theta}}_{ee}$ solution de l'équation (2.16).

2.4 Estimateurs assistés par un modèle

Lorsqu'on fait une enquête pour étudier des phénomènes socio-économiques, on dispose d'informations pré-existantes. On a déjà vu que cette information peut prendre la forme d'information auxiliaire. Mais on peut également tirer parti de connaissances sur les relations entre la variable d'intérêt et les variables auxiliaires. L'idéal serait bien sûr de connaître le modèle de superpopulation, mais on verra qu'un modèle de projection linéaire de la variable y sur le vecteur des variables auxiliaires $\boldsymbol{x}_k$ peut également être exploitable à l'étape de l'échantillonnage ou à l'étape de l'estimation dans une approche assistée par un modèle. On parle alors de modèle de travail ou de modèle d'ajustement qui peut être différent du modèle de superpopulation.

L'utilisation d'un modèle de travail sur la variable d'intérêt permet d'optimiser le choix du plan de sondage et/ou de l'estimateur dans le but d'obtenir des estimations plus précises. Dans cette section, on s'intéresse au choix d'un estimateur plus précis que l'estimateur par expansion.

Il faut toutefois rester vigilant car, lorsque le modèle de travail est différent du modèle de superpopulation, on s'expose au risque d'avoir un estimateur assisté par le modèle qui reste asymptotiquement sans biais, mais dont la variance est plus élevée que celle de l'estimateur par expansion. C'est un cas de figure qui peut se produire lorsqu'on utilise un modèle de régression linéaire sans constante alors que les données suivent un modèle de régression linéaire avec constante (voir Hansen, Madow et Tepping, 1983). On peut également citer les travaux d'Andersson et Thorburn (2005) qui mettent en évidence une perte de précision lors de l'utilisation d'un modèle de régression linéaire comme modèle de travail lorsque le plan de sondage est stratifié.

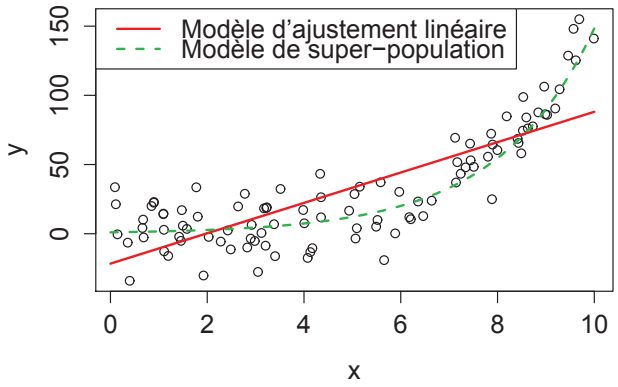


FIGURE 2.1 – Modèle d'ajustement linéaire sur la population $\mathcal{U}$

2.4.1 Estimateur assisté par un modèle de régression

Pour construire un estimateur de t_y assisté par le modèle, on utilise un modèle d'ajustement du type :

$$\forall k \in \mathcal{U}, \quad y_k = m(\boldsymbol{B}^T \boldsymbol{x}_k) + u_k, \tag{2.17}$$

où le vecteur de paramètre $\boldsymbol{B}$ est choisi de façon à minimiser

$$\sum_{k \in \mathcal{U}} \left\{ y_k - m(\boldsymbol{B}^T \boldsymbol{x}_k) \right\}^2$$

et le résidu vaut simplement par différence $u_k = y_k - m(\boldsymbol{B}^T \boldsymbol{x}_k)$.

Remarque 3. *Les résidus $u_k = y_k - m(\boldsymbol{B}^T \boldsymbol{x}_k)$ n'ont pas besoin d'être indépendants de $\boldsymbol{x}_k$ comme l'illustre la Figure (2.1) dans le cas d'un modèle d'ajustement linéaire en présence d'un modèle de superpopulation quadratique.*
Dans une approche basée sur le modèle (qu'on appelle également approche par modèle de prédiction et qu'on verra à la Section 2.6), il en va autrement ; on cherche un modèle de prédiction où le terme d'erreur est bien indépendant des covariables. Ce type de modèle est particulièrement important pour l'inférence en présence de non-réponse.

On s'appuie sur le modèle d'ajustement (2.17) pour réécrire notre paramètre de population finie t_y comme :

$$t_y = \sum_{k \in \mathcal{U}} m(\boldsymbol{B}^T \boldsymbol{x}_k) + \sum_{k \in \mathcal{U}} u_k,$$

ce qui mène à l'estimateur assisté par le modèle :

$$\begin{aligned}\hat{t}_{y,AM}^{(1)} &= \sum_{k\in\mathcal{U}} m\left(\boldsymbol{B}^T\boldsymbol{x}_k\right) + \sum_{k\in s} d_k u_k \\ &= \sum_{k\in\mathcal{U}} m\left(\boldsymbol{B}^T\boldsymbol{x}_k\right) + \hat{t}_{u,\pi} \\ &= \hat{t}_{y,\pi} + \left\{\sum_{k\in\mathcal{U}} m\left(\boldsymbol{B}^T\boldsymbol{x}_k\right) - \sum_{k\in s} d_k m\left(\boldsymbol{B}^T\boldsymbol{x}_k\right)\right\}.\end{aligned}$$

Si le vecteur de paramètres $\boldsymbol{B}$ n'est pas connu (ou fixé), il peut être défini sur la population $\mathcal{U}$, par exemple, par les équations estimantes :

$$\sum_{k\in\mathcal{U}} m'\left(\boldsymbol{B}^T\boldsymbol{x}_k\right)\left\{y_k - m\left(\boldsymbol{B}^T\boldsymbol{x}_k\right)\right\}\boldsymbol{x}_k = \boldsymbol{0},$$

et estimé à partir des données de l'échantillon par les équations estimantes

$$\sum_{k\in s} d_k\, m'\left(\hat{\boldsymbol{B}}_{ee}^T\boldsymbol{x}_k\right)\left\{y_k - m\left(\hat{\boldsymbol{B}}_{ee}^T\boldsymbol{x}_k\right)\right\}\boldsymbol{x}_k = \boldsymbol{0}.$$

On obtient alors un second estimateur assisté par le modèle de t_y qui peut être écrit sous une forme « projection » :

$$\hat{t}_{y,AM}^{(2)} = \sum_{k\in\mathcal{U}} m\left(\hat{\boldsymbol{B}}_{ee}^T\boldsymbol{x}_k\right) + \sum_{k\in s} d_k\tilde{u}_k$$

où $\tilde{u}_k = y_k - m\left(\hat{\boldsymbol{B}}_{ee}^T\boldsymbol{x}_k\right)$.
On peut également l'écrire sous la forme d'un estimateur par la différence :

$$\hat{t}_{y,AM}^{(2)} = \hat{t}_{y,\pi} + \left\{\sum_{k\in\mathcal{U}} m\left(\hat{\boldsymbol{B}}_{ee}^T\boldsymbol{x}_k\right) - \sum_{k\in s} d_k m\left(\hat{\boldsymbol{B}}_{ee}^T\boldsymbol{x}_k\right\}\right), \tag{2.18}$$

Lorsque la fonction $m(\cdot)$ est la fonction identité $m(x) = x$, alors nous obtenons l'estimateur classique GREG linéaire de Särndal et al. (1992).
Dans sa formulation générale, l'estimateur (2.18) constitue une généralisation de l'estimateur GREG linéaire. Lehtonen and Veijanen (1998) and Lehtonen et al. (2003, 2005) ont introduit des estimateurs assistés par des modèles de régression logistique, de régression logistique multinomiale et des modèles mixtes. Une classe d'estimateurs par la régression généralisée étendue a également été proposée par Montanari and Ranalli (2002). Ces approches fonctionnent bien lorsque la variable d'intérêt suit un modèle non linéaire, toutefois l'information auxiliaire doit être connue pour toutes les unités de la population $\mathcal{U}$.

2.4.2 Exemple de l'estimateur par la régression linéaire

Supposons que la variable y est liée au vecteur $\boldsymbol{x}$ suivant un modèle qui n'est pas connu avec précision. On utilise un modèle d'ajustement linéaire entre y and $\boldsymbol{x}$ qui correspond au modèle (2.17) où la fonction $m(\cdot)$ est la fonction identité :

$$\forall k \in \mathcal{U}, \quad y_k = \boldsymbol{B}^T\boldsymbol{x}_k + u_k. \tag{2.19}$$

On peut alors décomposer le total t_y de la façon suivante :

$$t_y \;=\; \boldsymbol{B}^T \boldsymbol{t}_x + t_u,$$

où $\boldsymbol{B}$ est défini par une méthode des moindres carrés pondérés par les poids c_k

$$\boldsymbol{B} \;=\; \left(\sum_{k\in\mathcal{U}} c_k^{-1} \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum_{k\in\mathcal{U}} c_k^{-1} \boldsymbol{x}_k y_k,$$

$\boldsymbol{t}_x = \sum_{k\in\mathcal{U}} \boldsymbol{x}_k$ et $t_u = \sum_{k\in\mathcal{U}} \left(y_k - \boldsymbol{B}^T \boldsymbol{x}_k\right)$.

On construit l'estimateur de t_y assisté par le modèle d'ajustement linéaire (Särndal et al. ,1992) :

$$\hat{t}_{y,AM} \;=\; \hat{\boldsymbol{B}}_\pi^T \boldsymbol{t}_x + \sum_{k\in s} d_k \left(y_k - \hat{\boldsymbol{B}}_\pi^T \boldsymbol{x}_k\right) \tag{2.20}$$

$$\;=\; \hat{t}_{y,\pi} + \hat{\boldsymbol{B}}_\pi^T \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,\pi}\right), \tag{2.21}$$

où

$$\hat{\boldsymbol{B}}_\pi \;=\; \left(\sum_{k\in s} c_k^{-1} d_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum_{k\in s} c_k^{-1} d_k \boldsymbol{x}_k y_k$$

Cet estimateur est un estimateur complexe. Sous certaines conditions, cet estimateur est convergent et son développement de Taylor peut s'écrire

$$N^{-1}\left(\hat{t}_{y,AM} - t_y\right) \;=\; N^{-1} \sum_{k\in s} d_k \left(y_k - \boldsymbol{B}^T \boldsymbol{x}_k\right) + O_P(n^{-1}).$$

Sa variance de Taylor vaut

$$\begin{aligned} A\mathbb{V}\left(\hat{t}_{y,AM}\right) \;&=\; \mathbb{V}_P\left(\sum_{k\in s} d_k \left(y_k - \boldsymbol{B}^T \boldsymbol{x}_k\right)\right) \\ &=\; \sum_{k\in\mathcal{U}} \sum_{l\in\mathcal{U}} \left(\pi_{k,l} - \pi_k \pi_l\right) d_k \left(y_k - \boldsymbol{B}^T \boldsymbol{x}_k\right)\; d_l \left(y_l - \boldsymbol{B}^T \boldsymbol{x}_l\right). \end{aligned}$$

Elle est estimée par

$$\widehat{A\mathbb{V}}\left(\hat{t}_{y,AM}\right) \;=\; \sum_{k\in s} \sum_{l\in s} \frac{\left(\pi_{k,l} - \pi_k \pi_l\right)}{\pi_{k,l}}\; d_k \left(y_k - \hat{\boldsymbol{B}}_\pi^T \boldsymbol{x}_k\right)\; d_l \left(y_l - \hat{\boldsymbol{B}}_\pi^T \boldsymbol{x}_l\right).$$

La variance de l'estimateur assisté par le modèle est plus faible que la variance de l'estimateur par expansion dans la mesure où le modèle de travail est relativement bien ajusté. Dans le cas où le modèle d'ajustement linéaire ne comporte pas de constante, on peut facilement créer un estimateur assisté par le modèle moins précis que l'estimateur par expansion.

On peut remarquer que si le plan de sondage est un tirage équilibré (Deville et Tillé (2004) ou Fuller(2009b)) avec les équations d'équilibrage $\hat{\boldsymbol{t}}_{x,\pi} = \boldsymbol{t}_x$, alors l'estimateur assisté par le modèle $\hat{t}_{y,AM}$ est égal à l'estimateur Horvitz-Thompson $\hat{t}_{y,\pi}$.

L'estimateur assisté par le modèle $\hat{t}_{y,AM}$ est connu sous le nom d'estimateur par la régression généralisé (estimateur GREG) et noté $\hat{t}_{y,GREG}$. Cet estimateur occupe une place importante en sondage et on peut consulter Fuller (2002) ou Särndal et al. (1992) pour avoir une description de cette classe des estimateurs par la régression. La linéarité du modèle d'ajustement présente l'avantage de nécessiter moins d'information auxiliaire qu'un modèle non linéaire. En effet, il suffit de connaître les totaux des variables auxiliaires et non les valeurs de $\boldsymbol{x}$ pour toutes les unités de la base de sondage.

Enfin, il faut noter que l'estimateur GREG peut s'écrire comme un estimateur linéaire pondéré. Toutefois les pseudo-poids

$$w_k = d_k \left\{ 1 + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,\pi}\right)^T \left(\sum_{k \in s} c_k^{-1} d_k \boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1} c_k^{-1} \boldsymbol{x}_k \right\}$$

ne sont pas forcément tous supérieurs à 0 (et a fortiori à 1), contrairement aux poids d'échantillonnages. La méthode du calage apportera une solution à cet inconvénient des poids négatifs.

2.5 Estimateurs par calage

Le calage est une méthode d'ajustement des poids d'un estimateur linéaire pondéré inital dans le but d'obtenir un nouvel estimateur linéaire qui estime parfaitement un certain nombre de totaux connues sur la population $\mathcal{U}$ (qu'on appellera totaux de calage ou totaux de contrôle). C'est une méthode qui permet d'intégrer de façon systématique l'information auxiliaire disponible.
Les poids obtenus par cette méthode sont appelés **poids de calage**.
La force et la popularité du calage auprès des praticiens reposent sur cette **cohérence** entre les estimations et les vraies valeurs des totaux de calage mais aussi sur le fait qu'il s'agit d'une méthode facile à mettre en œuvre pour prendre en compte l'information auxiliaire. Avec un estimateur par calage, on peut espérer que la « stratégie de sondage » a une meilleure représentativité que la stratégie avec l'estimateur initial. Toutefois, le gain (ou la perte) en précision de l'estimateur par calage dépend du modèle de superpopulation lié à la variable d'intérêt y. Nous reviendrons sur ce point lors du calcul de la variance.

La méthode du calage permet de construire une classe d'estimateurs pondérés linéaires, appelés **estimateurs par calage**. Cette classe d'estimateurs contient de nombreux estimateurs connus comme l'estimateur GREG, l'estimateur par le ratio ou l'estimateur post-stratifié.

Pour des tailles d'échantillon finies, les estimateurs par calage sont en général biaisés et leur variance n'est pas calculable. Si l'estimateur initial est convergent (comme l'estimateur par expansion), alors les estimateurs par calage associés sont également convergents (Kott, 2006). Pour des échantillons de grande taille (i.e. asymptotiquement), on montre que l'estimateur par calage est équivalent à un estimateur GREG (ou de type GREG comme nous le verrons à la fin de cette section).

Nous allons présenter la méthode du calage dans le cas classique où l'on cale sur des totaux $\boldsymbol{t}_x$ connus sur la population $\mathcal{U}$ et où l'estimateur initial est l'estimateur par expansion. Les estimateurs par calage sont de la forme

$$\hat{t}_{y,CAL} = \sum_{k \in s} w_k y_k,$$

où les poids de calage w_k, $k \in s$, vérifient les équations de calage

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \boldsymbol{t}_x, \tag{2.22}$$

où $\boldsymbol{x}$ est un vecteur de variables auxiliaires dont on connaît les valeurs pour l'échantillon s et le total $\boldsymbol{t}_x$.

2.5.1 Calage par minimisation de la distance entre les poids d'échantillonnage et les poids de calage (minimum distance method)

Deville et Särndal (1992), Deville et al. (1993) et Särndal (2007) ont proposé une classe d'estimateurs linéaires pondérés par des poids proches des poids d'échantillonnage et assurant la cohérence entre l'estimation et la vraie valeur du vecteur de totaux auxiliaires $\boldsymbol{t}_x$.

Les poids de calage sont obtenus par résolution du programme d'optimisation suivant :

$$\min_{\{w_k,\, k \in s\}} \sum_{k \in s} G_k(w_k, d_k)$$

sous les contraintes : $\sum_{k \in s} w_k \boldsymbol{x}_k = \boldsymbol{t}_x$, où $G_k(w, d)$ est une pseudo-distance qui mesure la distance entre les poids d'échantillonnage et les poids de calage (à la différence d'une distance, la pseudo-distance n'est pas nécessairement symétrique en ses deux arguments).

On suppose que la fonction $G_k(w, d)$ est dérivable par rapport à w, strictement convexe, avec une dérivée partielle $g_k(w, d) = \partial G_k(w, d)/\partial w$ continue et telle que $g_k(d, d) = 0$. Habituellement la fonction de distance est choisie de telle sorte que $g_k(w, d) = g(w/d)/q_k$, où q_k est une pondération « judicieusement choisie », $g(\cdot)$ une fonction continue, dérivable en 1, strictement croissante, telle que $g(1) = 0$ et $g^{'}(1) = 1$. On note $F(u) = g^{-1}(u)$ la fonction réciproque de $g(\cdot)$ que l'on nommera

fonction de calage par la suite. On suppose en outre que $F^{''}(0) < \infty$.

On résout le programme d'optimisation avec un lagrangien et on obtient les poids de calage

$$w_k = d_k F\left(q_k \hat{\boldsymbol{\lambda}}^\top \boldsymbol{x}_k\right),$$

où $\hat{\boldsymbol{\lambda}}$ est solution (s'il en existe) des équations de calage

$$\sum_{k \in s} d_k \boldsymbol{x}_k F\left(q_k \boldsymbol{\lambda}^\top \boldsymbol{x}_k\right) = \sum_{k \in \mathcal{U}} \boldsymbol{x}_k. \tag{2.23}$$

Afin d'assurer que le système (2.23) a une solution à partir d'une certaine taille $N \geq N_0$ de la population, on suppose que la matrice de variance covariance

$$\Sigma_{x,x} = \frac{1}{N} \sum_{k \in \mathcal{U}} q_k \boldsymbol{x}_k \boldsymbol{x}_k^{'}$$

est régulière.

On peut trouver de nombreux exemples de fonction de distance dans Deville et Särndal (1992). On citera ici, à titre d'exemple, trois distances qui correspondent à trois fonctions de calage particulièrement intéressantes.
Le première est la distance du χ^2 :

$$G_k(w_k, d_k) = (w_k - d_k)^2 / (2 q_k d_k).$$

Cette distance correspond à la fonction de calage dite linéaire $F(u) = 1 + u$ et les poids de calage obtenus sont appelés poids de calage linéaires et valent :

$$w_k = d_k \left(1 + q_k \boldsymbol{\lambda}^\top \boldsymbol{x}_k\right). \tag{2.24}$$

L'estimateur obtenu avec la fonction de calage linéaire est égal à l'estimateur GREG linéaire (2.20) présenté à la section précédente où les poids c_k valent q_k^{-1}.

Une seconde distance d'intérêt est

$$G_k(w_k, d_k) = w_k \log(w_k / d_k) - w_k + d_k$$

qui correspond à la fonction de calage exponentielle $F(u) = \exp(u)$. Les poids de calage exponentiels

$$w_k = d_k \exp(q_k \boldsymbol{\lambda}^\top \boldsymbol{x}_k) \tag{2.25}$$

ont l'avantage d'être positifs. En outre, la fonction de calage exponentielle est égale à sa fonction dérivée $h(u) = F^{'}(u) = \exp(u)$. L'intérêt de cette remarque apparaîtra ultérieurement dans le chapitre sur la correction de la non-réponse par calage.

Une troisième et dernière distance intéressante est la distance dite *logit* qui permet d'obtenir des poids bornés

$$w_k = d_k \frac{L(U-1) + U(1-L)\exp(A\ q_k \boldsymbol{\lambda}^\top \boldsymbol{x}_k)}{(U-1) + (1-L)\exp(A\ q_k \boldsymbol{\lambda}^\top \boldsymbol{x}_k)}, \tag{2.26}$$

où $A = (U-L)/\{(1-L)(U-1)\}$, et L et U sont les bornes inférieure et supérieure de la fonction de calage.

Deville et Särndal (1992) ont démontré un certain nombre de résultats concernant la convergence des estimateurs par calage et leur variance approchée.
D'abord, les poids de calage convergent vers les poids initiaux

$$w_k = d_k + O_P(n^{-0.5}).$$

Ensuite, les estimateurs par calage sont convergents par rapport au plan de sondage et

$$N^{-1}\left(\hat{t}_{y,CAL} - t_{y,N}\right) = O_P\left(n^{-0.5}\right),$$

où n est l'espérance de la taille de l'échantillon.
Enfin, sous les conditions mentionnées concernant la distance $G_k(\cdot,\ \cdot)$, les estimateurs par calage sont asymptotiquement équivalents à l'estimateur GREG :

$$N^{-1}\hat{t}_{y,CAL} = N^{-1}\hat{t}_{y,GREG} + O_p(n^{-1}). \tag{2.27}$$

En conséquence, on utilise la variance approchée de l'estimateur GREG pour construire un estimateur de variance des estimateurs par calage. On rappelle que la variance approchée de l'estimateur GREG vaut :

$$A\mathbb{V}(\hat{t}_{y,GREG}) \ = \ \sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}} (\pi_{k,l} - \pi_k\pi_l)\ d_k E_k\ d_l E_l,$$

où $E_k = y_k - \boldsymbol{B}^T\boldsymbol{x}_k$ et

$$\boldsymbol{B} \ = \ \left(\sum_{k\in\mathcal{U}} q_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum_{k\in\mathcal{U}} q_k \boldsymbol{x}_k y_k.$$

Les E_k sont les résidus de l'ajustement par les moindres carrés de y_k en fonction de $\boldsymbol{x}_k$ sur $\mathcal{U}$, pondérés par q_k. L'estimateur par calage sera donc d'autant plus précis que l'ajustement linéaire sera bon.

L'estimateur de variance proposé pour les estimateurs par calage est :

$$\hat{\mathbb{V}}(\hat{t}_{y,CAL}) \ = \ \sum_{k\in s}\sum_{l\in s} \frac{(\pi_{k,l} - \pi_k\pi_l)}{\pi_{k,l}}\ w_k e_k\ w_l e_l,$$

où $e_k = y_k - \hat{\boldsymbol{B}}_w^T\boldsymbol{x}_k$ et

$$\hat{\boldsymbol{B}}_w \ = \ \left(\sum_{k\in s} w_k q_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum_{k\in s} w_k q_k \boldsymbol{x}_k y_k.$$

2.5.2 Approche fonctionnelle du calage

Une approche alternative à la minimisation de la distance entre les poids d'échantillonnage d_k et les poids de calage w_k est l'approche dite par calage généralisé ou approche fonctionnelle développée par Deville (1998), Estevao et Särndal (2000, 2006) et Kott (2006). Cette approche élargit la famille des estimateurs par calage tout en intégrant les estimateurs dérivés des fonctions de distances présentées par Deville et Särndal.
On écrit les poids de calage directement sous forme d'une fonction de calage

$$w_k = d_k F(\boldsymbol{\lambda}^\top \boldsymbol{z}_k),$$

où $\boldsymbol{z}$ est un vecteur dont les valeurs sont connues pour les unités k de l'échantillon s et dont la dimension est égale à celle du vecteur $\boldsymbol{x}$, $F(0) = 0$ et $F^{'}(0) = 1$. Il n'est pas nécessaire de connaître le vecteur des totaux $\boldsymbol{t}_z$.
Le choix standard pour $\boldsymbol{z}_k$ est $\boldsymbol{z}_k = \boldsymbol{x}_k$ ou $\boldsymbol{z}_k = q_k \boldsymbol{x}_k$ pour des q_k donnés. Kott (2009) et Estevao et Särndal (2000) montrent toutefois qu'il est optimal (en termes de réduction de variance) de choisir :

$$\boldsymbol{z}_k = \sum_{l \in \mathcal{U}} (\pi_{k,l} - \pi_k \pi_l) \pi_k^{-1} (x_l / \pi_l). \tag{2.28}$$

L'estimateur par calage résultant est l'estimateur optimal de Montanari (1987). Dans le cas d'un sondage de Poison, $\boldsymbol{z}_k = (1 - \pi_k)\pi_k^{-1} x_k$. Toutefois, on verra que le choix d'un vecteur $\boldsymbol{z}_k$ différent du vecteur $\boldsymbol{x}_k$ trouve principalement son intérêt dans l'utilisation du calage pour corriger de la non-réponse où un biais conditionnel.

Le choix de la fonction de calage est guidé par les contraintes qu'on souhaite imposer aux poids de calage. En dehors de ce critère, il semble a priori que le choix de la fonction de calage a peu d'importance dans la mesure où toutes les fonctions donnent un estimateur asymptotiuement équivalent à l'estimateur GREG. Comme pour le choix du vecteur $\boldsymbol{z}_k$, on verra qu'une approche conditionnelle ou que le traitement de la non-réponse apporte une réponse à la question de la forme de la fonction de calage.

Comme le proposent Kim et Park (2010), on peut donner une expression encore plus générale des poids de calage dans l'approche fonctionnelle :

$$w_k = d_k F_k(\hat{\boldsymbol{\lambda}}).$$

Kim et Park (2010) montrent que dans ce contexte, l'estimateur par calage, utilisant les variables de calage $\boldsymbol{x}$ et les équations de calage

$$\sum_{k \in s} d_k F_k(\hat{\boldsymbol{\lambda}}) \boldsymbol{x}_k = \boldsymbol{t}_x,$$

peut s'écrire sous la forme d'un développement de Taylor :

$$\hat{t}_{y,CAL} = \hat{t}_{y,\pi} + \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x,\pi}\right)^\top \boldsymbol{B}_0 + o_p(Nn^{-0.5}),$$

où

$$\boldsymbol{B}_0 = \left\{ \sum_{k \in \mathcal{U}} \boldsymbol{h}_k(\boldsymbol{\lambda}_0) \boldsymbol{x}_k^T \right\}^{-1} \sum_{k \in \mathcal{U}} \boldsymbol{h}_k(\boldsymbol{\lambda}_0) y_k,$$

où $\boldsymbol{h}_k(\boldsymbol{\lambda}_0) = \partial F_k / \partial \boldsymbol{\lambda}(\boldsymbol{\lambda}_0)$ est la variable instrument et $\boldsymbol{\lambda}_0$ vérifie :

$$\sum_{k \in \mathcal{U}} F_k(\boldsymbol{\lambda}_0) \boldsymbol{x}_k = \boldsymbol{t}_x.$$

Kim et Park (2010) en déduisent un estimateur de la variance de l'estimateur par calage $\hat{t}_{y,CAL}$:

$$\hat{\mathbb{V}}(\hat{t}_{y,CAL}) = \sum_{k \in s} \sum_{l \in s} (\pi_{k,l} - \pi_k \pi_l) / \pi_{k,l} \; d_k \hat{F}_k \hat{e}_k \; d_l \hat{F}_l \hat{e}_l, \tag{2.29}$$

où

$$\hat{F}_k \hat{e}_k = F_k(\hat{\boldsymbol{\lambda}}) \left(y_k - \boldsymbol{x}_k^\top \hat{\boldsymbol{B}}_h \right)$$

et

$$\hat{\boldsymbol{B}}_h = \left\{ \sum_{k \in s} d_k \boldsymbol{h}_k(\hat{\boldsymbol{\lambda}}) \boldsymbol{x}_k^T \right\}^{-1} \sum_{k \in s} d_k \boldsymbol{h}_k(\hat{\boldsymbol{\lambda}}) y_k.$$

L'estimateur de variance (2.29) est convergent à la condition que :

$$\mathbb{V}_P \left(\hat{t}_{y,CAL} \right) = \mathbb{V}_P \left\{ \sum_{k \in s} d_k F_k(\boldsymbol{\lambda}_0)(y_k - \boldsymbol{B}_0^\top \boldsymbol{x}_k) \right\} + o(N^2 n^{-1}), \tag{2.30}$$

ce que les auteurs justifient par quelques conditions de régularités.

Remarque 4. *Le calage a été généralisé dans diverses directions. On peut citer les travaux de Wu and Sitter (2001) sur le model calibration qui définissent des équations de calage avec des variables de calage qui sont des prédictions des variables d'intérêt $\tilde{y}_k$ obtenues grâce à un modèle paramétrique non linéaire :*

$$\sum_{k \in s} w_k \tilde{y}_k = \sum_{k \in \mathcal{U}} \tilde{y}_k.$$

2.6 Estimateurs basés sur un modèle de prédiction

Nous avons vu à la Section 2.4 que l'utilisation d'un modèle d'ajustement sur la population $\mathcal{U}$ permettait de construire un estimateur assisté par le modèle dont l'inférence était toujours faite à partir du plan de sondage.

Dans cette section, on présente un cadre d'estimation et d'inférence d'un paramètre de recensement t_y basé sur un modèle de prédiction de y (modèle de superpopulation)

estimé à partir des données de l'échantillon s (voir Royall (1976, 1970) et Valliant et al. (2000)).
Pour fixer les idées, supposons que la variable d'intérêt y suive le modèle de régression linéaire suivant :

$$Y_k = \boldsymbol{\beta}^\top \boldsymbol{X}_k + \varepsilon_k, \ \forall k \in \mathcal{U} \tag{2.31}$$

où les ε_k sont des variables indépendantes, identiquement distribuées, d'espérance nulle $\mathbb{E}_m(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k) = 0$ et de variance $\mathbb{V}_m(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k) = \sigma^2$.

2.6.1 Plans de sondage ignorables

On suppose que le plan de sondage est ignorable (non-informatif) pour l'estimation du paramètre $\boldsymbol{\beta}$ (voir par exemple Skinner, Holt et Smith (1989), Pfeffermann (1993), Pfeffermann and Sverchkov (1999), Krieger et Pfeffermann (1992) et Valliant et al. (2000)). Un tel plan nécessite que la variable y ne soit pas utilisée pour sélectionner l'échantillon. Ainsi le plan de sondage n'est pas fonction des valeurs de la variable d'intérêt $(y_1, \ldots, y_N)$:

$$p(s) = f\left\{(\boldsymbol{x}_k, I_k;\ k \in \mathcal{U}),\ n,\ \boldsymbol{\Phi}\right\},$$

où $\boldsymbol{\Phi}$ est un vecteur de paramètres connus qui peut contenir, par exemple, la taille de la population.
La Figure (2.2) illustre les relations entre les caractéristiques des unités de la population $\mathcal{U}$ et le vecteur d'appartenance à l'échantillon $\boldsymbol{I} = (I_1, \ldots, I_N)^\top$. Le graphe indique que la variable d'intérêt représentée par le vecteur $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^\top$ est liée à la matrice $\boldsymbol{X}$ (dont les lignes sont les vecteurs $\boldsymbol{x}_k^\top$) mais n'est pas liée directement au vecteur $\boldsymbol{I}$. En d'autres termes, $\boldsymbol{Y}$ et $\boldsymbol{I}$ sont indépendants conditionnellement à $\boldsymbol{X}$.

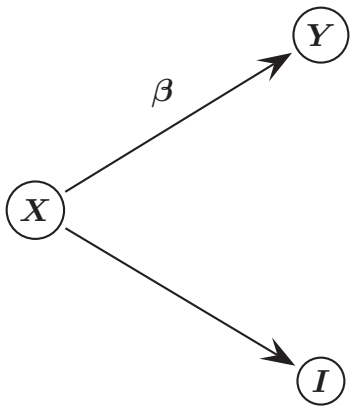


FIGURE 2.2 – Relations entre $\boldsymbol{X}$, $\boldsymbol{Y}$ et $\boldsymbol{I}$

Les plans de sondage probabilistes usuels sont ignorables. Ainsi, le plan de sondage aléatoire simple de taille n vaut $p(s) = \binom{N}{n} \mathbb{1}_{(\sum_{k \in \mathcal{U}} I_k = n)}$.

Pour un plan de sondage de Poison, à probabilités d'inclusion proportionnelles à la taille de la variable x, on a : $p(s) = \prod_{k \in \mathcal{U}} \left(\frac{n x_k}{t_x} \right)^{I_k}$.

Par contre, une enquête auprès des nouveaux nés, où la probabilité d'inclusion dépendrait du poids du bébé qui est la variable d'intérêt, ne correspond pas à un plan de sondage ignorable.

Lorsque le plan de sondage est ignorable, on peut montrer (Valliant et al., 2000) que le modèle statistique des $(Y_k;\ k \in s)$ est analogue au modèle (2.31) des $(Y_k;\ k \in \mathcal{U})$:

$$Y_k = \boldsymbol{\beta}^\top \boldsymbol{X}_k + \varepsilon_k,\ \forall k \in s$$

où les ε_k sont des variables indépendantes, identiquement distribuées, d'espérance nulle $\mathbb{E}_m \left(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k \right) = 0$ et de variance $\mathbb{V}_m \left(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k \right) = \sigma^2$.

A partir de l'échantillon s, on peut alors procéder à l'estimation du vecteur de paramètres $\boldsymbol{\beta}$ avec la méthode d'estimation qu'on utiliserait pour les données observées sur la population $\mathcal{U}$.

2.6.2 Meilleur prédicteur linéaire sans biais de t_y

On note $\hat{\boldsymbol{\beta}}_s$ l'estimateur sans biais de variance minimum du vecteur de paramètres $\boldsymbol{\beta}$ du modèle (2.31) à partir des données de l'échantillon s

$$\hat{\boldsymbol{\beta}}_s = \left\{ \sum_{k \in s} \boldsymbol{x}_k \boldsymbol{x}_k^\top \right\}^{-1} \sum_{k \in s} \boldsymbol{x}_k y_k. \tag{2.32}$$

La variance de $\hat{\boldsymbol{\beta}}_s$ (par rapport au modèle) vaut

$$\mathbb{V}_m \left\{ \hat{\boldsymbol{\beta}}_s \mid (\boldsymbol{X}_l = \boldsymbol{x}_l, l \in \mathcal{U}) \right\} = \sigma^2 \left(\sum_{k \in s} \boldsymbol{x}_k \boldsymbol{x}_k^\top \right)^{-1}.$$

Dans l'approche par prédiction, le meilleur prédicteur linéaire sans biais sous le modèle de t_y est (Valliant et al., 2000) :

$$\hat{t}_{y,m} = \sum_{k \in s} \left\{ 1 + \left(\sum_{k \notin s} \boldsymbol{x}_k \right) \left(\sum_{k \in s} \boldsymbol{x}_k \boldsymbol{x}_k^\top \right)^{-1} \boldsymbol{x}_k \right\} y_k. \tag{2.33}$$

L'estimateur (2.33) peut s'écrire également sous la forme

$$\hat{t}_{y,m} = \sum_{k \in s} y_k + \sum_{k \notin s} \hat{\boldsymbol{\beta}}_s^\top \boldsymbol{x}_k$$

qui offre une meilleure compréhension de la logique de l'approche par prédiction.

Cet estimateur est sans-biais sous le modèle, c'est-à-dire qu'il vérifie

$$\mathbb{E}_m \left\{ \hat{t}_{y,m} - t_y \mid (\boldsymbol{X}_l = \boldsymbol{x}_l, l \in \mathcal{U}) \right\} = 0,$$

et on peut montrer que la variance sous le modèle vaut

$$\begin{aligned} \mathbb{E}_m \left\{ \left(\hat{t}_{y,m} - t_y \right)^2 \mid (\boldsymbol{X}_l = \boldsymbol{x}_l, l \in \mathcal{U}) \right\} &= N^2 \left(1 - \frac{n}{N} \right) \frac{\sigma^2}{n} \\ &\times \left\{ \frac{n}{N} + \frac{n}{N(N-n)} \left(\sum_{k \notin s} \boldsymbol{x}_k \right)^\top \left(\sum_{k \in s} \boldsymbol{x}_k \boldsymbol{x}_k^\top \right)^{-1} \left(\sum_{k \notin s} \boldsymbol{x}_k \right) \right\}. \end{aligned}$$

Valliant et al. (2000, P.29-30) donnent une expression pour des modèles linéaires plus généraux.
Ces estimateurs par prédiction sont souvent plus efficaces que les estimateurs assistés par un modèle. Par ailleurs, ils permettent l'utilisation des techniques statistiques classiques. Il faut bien noter que lorsqu'on utilise une approche basée sur un modèle pour construire l'estimateur de t_y, on utilise également une inférence par rapport au modèle. On raisonne donc conditionnellement à l'échantillon tiré.

2.6.3 Approche par prédiction et robustesse

Les estimateurs par prédiction sont peu robustes aux mauvaises spécifications du modèle de prédiction. Pour apporter de la robustesse à ces estimateurs, on peut chercher un plan de sondage dont les probabilités d'inclusion sont égales aux inverses des pondérations de l'estimateur linéaire par prédiction. On aura alors un estimateur par expansion (approche par randomisation) égal à l'estimateur par prédiction. En cas de mauvaise spécification du modèle, l'estimateur profite d'une protection apportée par l'inférence à partir du plan de sondage.

2.6.3.1 Exemple du modèle d'analyse de la variance

On suppose que pour chaque strate $\mathcal{U}_h$, $h \in (1, \ldots, H)$,

$$Y_k = \mu_h + \varepsilon_k, \ \forall k \in \mathcal{U}$$

où les ε_k sont des variables indépendantes, identiquement distribuées, d'espérance nulle $\mathbb{E}_m \left(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k \right) = 0$ et de variance $\mathbb{V}_m \left(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k \right) = \sigma_h^2$.
L'estimateur par prédiction vaut

$$\hat{t}_{y,m} = \sum_h \sum_{k \in s_h} \frac{N_h}{n_h} y_k,$$

où N_h est la taille de la strate $\mathcal{U}_h$ et n_h la taille de l'échantillon s_h dans la strate $\mathcal{U}_h$.
Si on utilise un plan de sondage aléatoire stratifié, on obtient un estimateur par expansion égal à l'estimateur par la prédiction.

2.6.3.2 Exemple du modèle linéaire affine (Ratio)

On suppose que

$$Y_k = RX_k + X_k\varepsilon_k,\ \forall k \in \mathcal{U}$$

où les ε_k sont des variables indépendantes, identiquement distribuées, d'espérance nulle $\mathbb{E}_m\left(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k\right) = 0$ et de variance $\mathbb{V}_m\left(\varepsilon_k \mid \boldsymbol{X}_k = \boldsymbol{x}_k\right) = \sigma^2$.
L'estimateur par prédiction vaut

$$\hat{t}_{y,m} = \sum_{k \in s} \frac{N}{n} \frac{\bar{X}}{\bar{X}_s} y_k,$$

où $\bar{X}$ est la moyenne de la variable x sur la population $\mathcal{U}$ et $\bar{X}_s$ est la moyenne de la variable x sur l'échantillon.
Si on utilise un plan de sondage à probabilités d'inclusion égales, équilibré sur le total de la variable x, alors on obtient une fois encore l'égalité des estimateurs par expansion et par prédiction.

2.7 La non-réponse

Lors d'une enquête, un certain nombre d'unités interrogées refuse de répondre totalement ou partiellement. La non-réponse totale correspond aux unités pour lesquelles aucune donnée n'est collectée ; soit parce qu'elles ne sont pas joignables soit parce qu'elles refusent de répondre. La non-réponse partielle concerne les unités qui répondent à une partie seulement des questions posées. La non-réponse engendre des valeurs manquantes dans les fichiers de données qui empêchent l'utilisation directe des estimateurs habituels, fondés sur un fichier de données complet. Deux grandes familles de méthode existent pour traiter la non-réponse : la repondération des unités répondantes et l'imputation des valeurs manquantes. En général, la non-réponse totale est traitée par repondération et la non-réponse partielle par imputation.

On examine ici le cas de la non-réponse totale et on considère qu'il n'y a pas de non-réponse partielle. Une unité k est donc soit répondante, soit non-répondante. L'estimation du total t_y sera faite à partir des unités répondantes et avec des poids ajustés pour compenser la perte des non-répondants.

En présence de non-réponse, on peut être tenté d'utiliser des estimateurs prévus pour des données sans non-réponse en calculant simplement des moyennes sur l'échantillon des répondants au lieu de l'échantillon total s. Toutefois, lorsque le mécanisme de réponse est lié à la variable d'intérêt, ces estimateurs conçus pour des données complètes sont biaisés par rapport au modèle de non-réponse. Par exemple, dans une enquête sur le patrimoine, si les hauts patrimoines ont des taux de réponse plus faibles que la moyenne alors l'estimation de la moyenne des patrimoines sera sous-évaluée.

Prenons l'exemple d'un mécanisme de réponse, supposé indépendant du plan de sondage, qui corresponde à un plan de sondage de Poisson où les probabilités d'inclusion sont les probabilités de réponse p_k . On définit les indicatrices de réponse R_k qui valent 1 si l'unité k répond et 0 sinon. L'espérance par rapport à ce mécanisme de sélection est notée $\mathbb{E}_q(\cdot)$.

L'estimateur naïf qui utilise la moyenne sur l'échantillon des répondants vaut

$$\hat{t}_{y,\text{ naïf}} = \frac{\sum_{k\in s} d_k R_k y_k}{\sum_{k\in s} d_k R_k} N.$$

On peut montrer à l'aide d'un développement de Taylor (Lesage et Haziza, 2013)que le biais approché (par rapport au plan et au modèle de non-réponse) vaut

$$B_{P,q}(\hat{t}_{y,\text{ naïf}}) \approx \frac{1}{\bar{p}} \left\{ \sum_{k\in\mathcal{U}} (p_k - \bar{p})\, y_k \right\}, \tag{2.34}$$

où $\bar{p} = \sum_{k\in\mathcal{U}} p_k/N$ est le taux de réponse moyen. Si la covariance empirique entre y et p est non nulle alors l'estimateur naïf est biaisé. Si le taux de réponse moyen est faible ce biais est amplifié.

On considère un modèle de superpopulation sur les variables y et $\boldsymbol{x}$ et on combine les trois modèles : modèle de superpopulation, plan de sondage et modèle de non-réponse. On note $\mathbb{E}(\cdot) = \mathbb{E}_m \left[\mathbb{E}_P \left\{\mathbb{E}_q(\cdot)\right\}\right]$ l'espérance par rapport aux trois modèles combinés.
Le biais (2.34) tend asymptotiquement, sous le modèle combiné, vers :

$$N \frac{\mathbb{C}ov\left(R_k,\, Y_k\right)}{\mathbb{E}\left(R_k\right)}. \tag{2.35}$$

On en déduit qu'une covariance nulle entre Y_k et R_k nous permet d'obtenir un estimateur approximativement sans biais. Une autre condition favorable mais plus forte serait que R_k et Y_k soient indépendants. Dans ce cas, la non-réponse peut être ignorée puisqu'elle ne déforme pas la distribution de la variable d'intérêt y. On parle de non-réponse MCAR (Missing Completely At Random). La non-réponse n'a alors qu'un impact sur la taille de l'échantillon qu'elle réduit, ce qui entraine une augmentation de variance.

2.7.1 La non-réponse est MAR conditionnellement à des variables auxiliaires $\boldsymbol{X}_k$

D'autres cas de non-réponse peuvent être traités. Citons le cas où les covariables communes au modèle de superpopulation sur y et au modèle de réponse sont des variables auxiliaires (voir la Figure (2.3)). On parle alors d'une non-réponse MAR (Missing At Random) pour la variable y conditionnellement aux variables auxiliaires $\boldsymbol{X}_k$.
On peut définir ce cas par des conditions plus ou moins fortes d'indépendance entre y et R :

1. R_k et Y_k sont indépendantes conditionnellement à $\boldsymbol{X}_k$,
2. $\mathbb{E}(R_k \mid Y_k, X_k) = \mathbb{E}(R_k \mid X_k)$,
3. $\mathbb{E}(R_k Y_k \mid X_k) = \mathbb{E}(R_k \mid X_k)\mathbb{E}(Y_k \mid X_k)$,
4. $\mathbb{E}(Y_k \mid X_k, R_k = 1) = \mathbb{E}(Y_k \mid X_k)$,

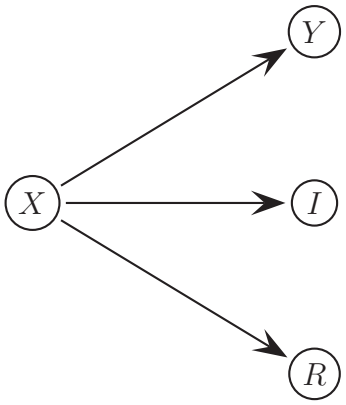


FIGURE 2.3 – Relation entre les variables Y, X, I et R

Proposition 1. *Les conditions (3) et (4) sont équivalentes. La condition (1) implique la condition (2) qui implique la condition (3).*

Démonstration. 1. La condition (1) implique la condition (2) par définition de l'indépendance conditionnelle.

2. Par conditionnement on a :

$$\begin{aligned} \mathbb{E}(R_k Y_k \mid X_k) &= \mathbb{E}\left\{\mathbb{E}(R_k Y_k \mid X_k, Y_k) \mid X_k\right\} \\ &= \mathbb{E}\left\{Y_k \mathbb{E}(R_k \mid X_k, Y_k) \mid X_k\right\}. \end{aligned}$$

En utilisant la condition (2), on obtient alors :

$$\begin{aligned} \mathbb{E}(R_k Y_k \mid X_k) &= \mathbb{E}\left\{Y_k \mathbb{E}(R_k \mid X_k) \mid X_k\right\} \\ &= \mathbb{E}(R_k \mid X_k)\mathbb{E}\left\{Y_k \mid X_k\right\}, \end{aligned}$$

qui correspond à la condition (3).

3. En utilisant les événements complémentaires $[R_k = 0]$ et $[R_k = 1]$, on a :

$$\mathbb{E}(R_k Y_k \mid X_k) = \mathbb{E}(Y_k \mid X_k, R_k = 1)\mathbb{E}(R_k \mid X_k).$$

D'où

$$\mathbb{E}(Y_k \mid X_k, R_k = 1) = \mathbb{E}(Y_k \mid X_k) \iff \mathbb{E}(R_k Y_k \mid X_k) = \mathbb{E}(Y_k \mid X_k)\mathbb{E}(R_k \mid X_k),$$

c'est-à-dire que les conditions (3) et (4) sont équivalentes.

□

Pour cette section, on retient la condition (2), i.e.,

$$\mathbb{E}(R_k \mid Y_k, \boldsymbol{X}_k) = \mathbb{E}(R_k \mid \boldsymbol{X}_k) \tag{2.36}$$

comme définition de la non-réponse MAR conditionnellement à $\boldsymbol{X}_k$ pour l'estimation de t_y. En outre, on fait l'hypothèse que le plan de sondage et le mécanisme de non-réponse sont indépendants (i.e. indépendance des indicatrices I_k et R_k conditionnellement aux variables auxiliaires).

2.7.2 Le modèle de non-réponse : estimation de la probabilité de réponse

Les indricatrices de réponse R_k suivent des lois de Bernoulli indépendantes (conditionnellement à la population finie) de paramètre $p_k > 0$. Dans le cas MAR, d'après la condition (2.36), on obtient

$$p_k = \mathbb{E}(R_k \mid Y_k, \boldsymbol{X}_k) = \mathbb{E}(R_k \mid \boldsymbol{X}_k). \tag{2.37}$$

On note $p_k = h(\boldsymbol{x}_k; \gamma)$ où les $\boldsymbol{x}_k$ sont connus pour toutes les unités $k \in s$.

On peut estimer par modélisation paramétrique ce modèle de réponse et calculer des estimations des probabilités de réponse p_k, pour $k \in s$, notées $\hat{p}_k$.
Ces estimations $\hat{p}_k$ seront utilisées pour construire des estimateurs approximativement sans biais de t_y.

2.7.3 Le modèle de y en fonction de $\boldsymbol{x}$

Dans le cas MAR, d'après (2.36) et la proposition (1), on a $\mathbb{E}(Y_k \mid X_k, R_k = 1) = \mathbb{E}(Y_k \mid X_k)$ ce qui signifie que le modèle de y en fonction de $\boldsymbol{x}$ peut être estimé à partir de l'échantillon des répondants (paramétriquement ou non-paramétriquement).

Avec un modèle paramétrique du type $\mathbb{E}(Y_k \mid X_k) = f(\boldsymbol{\beta}; \boldsymbol{X}_k)$, un estimateur sans biais $\hat{\boldsymbol{\beta}}_r$ du vecteur de paramètres $\boldsymbol{\beta}$ est obtenu en résolvant le système d'équations estimantes sur les répondants :

$$\sum_{k \in s} d_k R_k \left\{ y_k - f\left(\hat{\boldsymbol{\beta}}_r; \boldsymbol{x}_k\right) \right\} \boldsymbol{x}_k = \mathbf{0}. \tag{2.38}$$

2.7.4 Estimateurs corrigés pour la non-réponse

Dans ce contexte de non-reponse MAR conditionnellement à $\boldsymbol{x}$, on peut penser utiliser au moins 3 estimateurs.

1. Un estimateur par approche « quasi-randomisation ».
   Dans cette approche, on modélise la non-réponse comme une ultime phase d'échantillonnage avec un plan de sondage de Poison de taille fixe où les probabilités d'inclusion sont les estimateurs des probabilités de réponse $\hat{p}_k$.

$$\hat{t}_y^{(1)} = \sum_{k \in s_r} d_k \frac{1}{\hat{p}_k} y_k,$$

2. Un estimateur par approche quasi-randomisation assisté par un modèle.
   Il s'agit d'utiliser une approche assistée par le modèle sur y :

$$\begin{aligned}\hat{t}_y^{(2)} &= \sum_{k\in s} d_k f\left(\hat{\boldsymbol{\beta}}; \boldsymbol{x}_k\right) + \sum_{k\in s_r} d_k \frac{y_k - f\left(\hat{\boldsymbol{\beta}}; \boldsymbol{x}_k\right)}{\hat{p}_k} \\ &= \sum_{k\in s_r} d_k \frac{1}{\hat{p}_k} y_k + \left\{\sum_{k\in s} d_k f\left(\hat{\boldsymbol{\beta}}; \boldsymbol{x}_k\right) - \sum_{k\in s_r} d_k \frac{f\left(\hat{\boldsymbol{\beta}}; \boldsymbol{x}_k\right)}{\hat{p}_k}\right\},\end{aligned}$$

3. Un estimateur par approche basée sur le modèle.
   Cet estimateur est en fait un estimateur par imputation.

$$\hat{t}_y^{(3)} = \sum_{k\in s_r} d_k y_k + \sum_{k\in s\setminus s_r} d_k f\left(\hat{\boldsymbol{\beta}}; \boldsymbol{x}_k\right),$$

   où $k \in s \setminus s_r$ correspond aux unités de s qui ne répondent pas à l'enquête.

2.7.5 Méthodes des groupes homogènes de non-réponse pour estimer les probabilités de réponses p_k

Une méthode classique d'estimation des probabilités de réponses p_k repose sur l'utilisation d'un modèle de régression logistique (Little, 1986) de R_k en fonction de $\boldsymbol{x}_k$ pour partitionner la population en groupes homogènes de non-réponse dans lesquels les prédictions de probabilités de réponse du modèle logistique sont proches. La probabilité de réponse d'un individu k est finalement estimée par le taux de réponse moyen du groupe auquel appartient l'unité k (on peut éventuellement pondérer par les poids de sondage).
Cette méthode permet de réduire la dispersion du facteur d'ajustement pour la non-réponse $\dfrac{1}{\hat{p}_k}$.

2.7.6 Correction de la non-réponse par calage et calage généralisé

Chang et Kott (2008), Deville (1998, 2002), Dupont (1996), Kott (2006, 2008), Lundström et Särndal (1999), Särndal et Lundström (2005, 2010) et Sautory (2003) ont proposé d'utiliser des estimateurs par calage pour corriger la non-réponse.
Dans la grande majorité des cas, ces auteurs traitent le mécanisme de réponse comme une ultime phase de sondage et le ratio des poids de calage sur les poids d'échantillonnage visent à estimer les inverses des probabilités de réponse p_k. Nous étudierons dans les chapitres 4 et 5 dans quelle mesure on peut effectivement utiliser des estimateurs par calage pour traiter la non-réponse.

2.8 L'inférence conditionnelle en sondage

2.8.1 L'inférence conditionnelle en sondage

Dans cette Section, nous abordons la question de l'inférence conditionnelle et de son corollaire : les estimateurs conditionnellement sans biais. Rao (1985) explique clairement pourquoi il est préférable de recourir à une inférence conditionnelle sous le plan de sondage au moment de l'estimation.

> In the conventional set-up for inference in survey sampling the sample design defines the sample space $\mathcal{S}$ (set of possible samples s) and the associated probabilities of selection, $P(s)$. The choice of an estimator is based on the criterion of consistency or unbiasedness and on the comparison of mean square errors (MSE), under repeated sampling with probabilities $P(s)$, using the sample space $\mathcal{S}$ as the reference set. [...]
> The comparison of unconditional mean square errors is appropriate at the design stage, but the sample space $\mathcal{S}$ **may not be the relevant reference set** for inference after the sample s has been drawn, if the sample contains "recognizable subsets". [...]

Ceci veut dire qu'il peut être plus approprié d'utiliser une loi de probabilité conditionnelle de l'estimateur du total t_y pour calculer l'espérance, la variance et l'intervalle de confiance.
Ainsi, s'il existe un lien entre les estimateurs $\hat{t}_{y,\pi}$ et $\hat{t}_{x,\pi}$, il est alors préférable d'utiliser la distribution de $\hat{t}_{y,\pi}$ conditionnelle à $\hat{t}_{x,\pi}$ plutôt que sa distribution marginale (ce qui est fait habituellement en utilisant $P(s)$).

La statistique ancillaire
La statistique choisie pour le conditionnement est une statistique ancillaire (Rao, 1985), i.e. une statistique dont la distribution est connue et qui ne dépend pas du paramètre à mesurer. Rao (1985) donne plusieurs exemples de statistique ancillaires reposant sur des tailles d'échantillon. Tillé (1998) utilise comme statistique ancillaire l'estimateur par expansion $\hat{t}_{x,\pi}$ du total t_x de la variable auxiliaire complète x (x_k connu pour $k \in \mathcal{U}$). Les valeurs prises par cette statistique sont censées permettre une partition de l'espace $\mathcal{S}$ en sous-ensembles plus homogènes, i.e. des sous-ensembles sur lesquels l'estimateur $\hat{t}_{y,\pi}$ prend des valeurs différentes. Ainsi, si on note s_0 l'échantillon sélectionné et t la valeur prise par $\hat{t}_{x,\pi}$ pour cet échantillon, alors l'inférence se fait sur le sous-ensemble $\mathcal{S}^{|t} = \left\{ s \mid \hat{t}_{x,\pi}(s) = t \right\}$.

La loi conditionnelle de $\hat{t}_{y,\pi}$ est construite à partir de la probabilité d'échantillonnage conditionnelle $P^{|t}(s) = P^{|\hat{t}_{x,\pi}=t}(s) = P\left(s \mid [\hat{t}_{x,\pi} = t]\right)$ qui vaut :

$$\begin{cases} P^{|t}(s) = 0 & \text{si } \hat{t}_{x,\pi}(s) \neq t \\ P^{|t}(s) = \dfrac{P(s)}{P([\hat{t}_{x,\pi}(s) = t])} & \text{si } \hat{t}_{x,\pi}(s) = t \end{cases} \tag{2.39}$$

On note $\mathbb{E}_{P|t}(\cdot)$ l'espérance conditionnelle : $\mathbb{E}_{P|t}(\cdot) = \mathbb{E}_P(\cdot \mid [\hat{t}_{x,\pi} = t])$. La variance conditionnelle est notée : $\mathbb{V}_{P|t}(\cdot)$.

On définit la probabilité d'inclusion conditionnelle $\pi_k^{|t}$ de l'unité k par

$$\pi_k^{|t} = \mathbb{E}_{P|t}(I_k)$$

et la probabilité d'inclusion jointe conditionnelle $\pi_{k,l}^{|t}$ des unités k et l par

$$\pi_{k,l}^{|t} = \mathbb{E}_{P|t}(I_k I_l).$$

Le biais conditionnel de l'estimateur par expansion vaut

$$BC_{P|t}(\hat{t}_{y,\pi}) = \mathbb{E}_{P|t}(\hat{t}_{y,\pi} \mid [\hat{t}_{x,\pi} = t]) - t_y. \tag{2.40}$$

Ce biais conditionnel peut être très important lorsque les variables y et x sont liées ou lorsque le paramètre est d'ordre $\alpha \neq 0$ avec un plan de sondage non équilibré sur la taille de la population. Royall et Cumberland (1981) ont montré l'existence de tels biais pour l'estimateur par le ratio.

Donnons un exemple simple pour illustrer le biais conditionnel et l'intérêt de l'approche conditionnnelle. On considère une population de taille N et un tirage bernoullien de taille d'échantillon $n(s)$ aléatoire telle que $\mathbb{E}_P\{n(s)\} = n$. L'échantillon s_0 sélectionné est de taille n_0. On peut montrer que

$$BC_{P|t}\left(\hat{t}_{y,\pi}\right) = \mathbb{E}_{P|t}(\hat{t}_{y,\pi} \mid [n(s) = n_0]) - t_y = \left\{\frac{n_0}{n} - 1\right\} t_y \tag{2.41}$$

et

$$\mathbb{V}_P\left(\hat{t}_{y,\pi} \mid [n(s) = n_0]\right) = \left(\frac{n_0}{n}\right)^2 N^2 \left(1 - \frac{n_0}{N}\right) \frac{S_y^2}{n_0} \tag{2.42}$$

La formule (2.41) montre que si $t_y \neq 0$ et $n_0 \neq n$, alors l'estimateur par expansion est conditionnellement biaisé et que ce biais peut être très grand puisque le biais relatif vaut $\dfrac{n_0 - n}{n}$. Ainsi on obtient un biais de 100% lorsque la taille de l'échantillon n_0 est le double de la taille espérée n.

Pour ce qui concerne la variance conditionnelle, il est plus probable qu'elle soit inférieure à la variance non-conditionnelle, mais ce n'est pas obligatoire. La variance non-conditionnelle est constante quelque soit la valeur n_0 et vaut

$$\begin{aligned}
\mathbb{V}_P\left(\hat{t}_{y,\pi}\right) &= N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{\sum_{k\in s} y_k^2}{N} \\
&= N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \left\{ \frac{N-1}{N} S_y^2 + \left(\frac{\sum_{k\in s} y_k}{N}\right)^2 \right\}.
\end{aligned}$$

Dans le cadre de l'inférence conditionnelle (conditionnement sur la taille de l'échantillon tiré $n(s)$), l'intervalle de confiance habituel $IC_{95\%}(\hat{t}_{y,\pi})$ n'est plus adapté car sous la loi conditionnelle

$$P^{|t}\left([t_y \in IC_{95\%}(\hat{t}_{y,\pi})]\right) \neq 95\%.$$

Dit autrement, la statistique ancillaire met en évidence que l'intervalle de confiance couvre mal la zone où le paramètre t_y a de forte chance de se trouver.

L'espace des échantillons conditionnel
On a déjà mentionné que le sous-ensemble des échantillons sur lequel on infère est :

$$\begin{aligned} \mathcal{S}^{|t} &= \left\{ s \mid \hat{t}_{x,\pi}(s) = t \right\} \\ &= \left\{ s \mid P^{|t}(s) > 0 \right\}. \end{aligned}$$

En pratique, cet espace peut être vide ou conduire à un trop petit nombre d'échantillons, on verra au Chapitre 6 qu'il est préférable de définir l'espace d'inférence $\mathcal{S}^{|t}$ de façon plus large :

$$\begin{aligned} \mathcal{S}_h^{|t} &= \left\{ s \mid t - h^{|t} \leqslant \hat{t}_{x,\pi}(s) \leqslant t + h^{|t} \right\} \\ &= \left\{ s \mid \Phi_{h|t}\left(\hat{t}_{x,\pi}\right) = 1 \right\} \end{aligned}$$

où $h^{|t} > 0$ est un paramètre qui peut être fixé par exemple de manière à obtenir des probabilités d'inclusion conditionnelles strictement positives et

$$\Phi_{h|t}\left(\hat{t}_{x,\pi}\right) = \mathbb{1}_{\left(-h^{|t} \leqslant \hat{t}_{x,\pi}(s) - t \leqslant h^{|t}\right)}.$$

2.8.2 Les estimateurs conditionnellement sans biais

Royal et Cumberland (1981) ont mis en évidence les risques de biais conditionnels de l'estimateur par le ratio lorsque le modèle suivi par les données s'éloignaient d'un modèle linéaire affine. Ils concluaient que la loi de probabilité du plan de sondage, $P(s)$, n'était pas nécessairement appropriée pour l'inférence. Dans leur exemple ils montrent que l'estimateur par le ratio de t_y et l'estimateur par expansion de t_x sont liés et qu'en conséquence connaissant $\hat{t}_{x,\pi}$, on peut mettre en évidence un biais conditionnel de l'estimateur par le ratio de t_y. La solution préconisée par Royall et Cumberland (1981) pour éviter le biais conditionnel de l'estimateur par le ratio était d'opter pour un **tirage équilibré** tel que $\hat{t}_{x,\pi} = t_x$.

Estimateur initial corrigé du biais conditionnel
Robinson (1987) a repris l'exemple de Royal et Cumberland (1981) et a estimé, sous certaines hypothèses, la valeur du biais conditionnel de l'estimateur par le ratio de t_y. A partir de cette estimation du biais, il a construit un nouvel estimateur qui était l'estimateur par le ratio diminué du biais conditionnel estimé.

Tillé (1999) développe l'approche de Robinson et montre qu'en retranchant à un estimateur HT une estimation de son biais conditionnel, on obtient un estimateur équivalent à l'estimateur linéaire optimal de Montanari (1987). Ce résultat contient notamment le cas de l'estimateur post-stratifié.

2.8.3 L'estimateur par expansion conditionnel

Cette recherche d'estimateur conditionnellement sans biais mène naturellement à la construction d'un estimateur par expansion conditionnel. Il s'agit de reprendre la démarche suivie par Horvitz et Thompson (1952) : chercher un estimateur linéaire pondéré conditionnellement sans biais. On cherche des poids w_k tels que pout toute variable d'intérêt y :

$$\mathbb{E}_P\left(\sum_{k\in\mathcal{U}} w_k I_k y_k \;\mid\; \Phi_{h|t}\left(\hat{t}_{x,\pi}\right)=1\right)=t_y.$$

Si $\pi_k^{|t}>0$, $\forall k\in\mathcal{U}$, Il suffit de choisir

$$\begin{aligned} w_k^{-1} &= \mathbb{E}\left(I_k \;\mid\; \Phi_{h|t}\left(\hat{t}_{x,\pi}\right)=1\right) && (2.43)\\ &= \pi_k^{|t}. && (2.44)\end{aligned}$$

Si $\pi_k^{|t}$ est nul pour certaines unités $k\notin s$, alors l'estimateur ne sera pas exactement conditionnellement sans biais. Le biais vaudra $\sum_{(k\in\mathcal{U},\,\pi_k|t=0)} y_k$. Il faut donc veiller à garder un ensemble d'inférence $\mathcal{S}^{|t}$ assez grand afin d'éviter les probabilités d'inclusion conditionnelles nulles.
Tillé (1998) a présenté cet estimateur sous le nom de l'estimateur simple conditionnellement pondéré (SCP).

La variance conditionnelle de cet estimateur vaut :

$$\mathbb{V}_{P|t}\left(\hat{t}_{y,\pi|t}\right) = \sum_{k,l\in\mathcal{U}}(\pi_{k,l}^{|t}-\pi_k^{|t}\pi_l^{|t})\frac{y_k}{\pi_k^{|t}}\frac{y_l}{\pi_l^{|t}}.$$

Un estimateur conditionnellement sans biais de cette variance conditionnelle est donnée par :

$$\hat{\mathbb{V}}_{P|t}\left(\hat{t}_{y,\pi|t}\right)=\sum_{k,l\in s}(\pi_{k,l}^{|t}-\pi_k^{|t}\pi_l^{|t})\frac{1}{\pi_{k,l}^{|t}}\frac{y_k}{\pi_k^{|t}}\frac{y_l}{\pi_l^{|t}},$$

lorsque les probabilités d'inclusion jointes $\pi_{k,l}^{|t}$ sont non nulles.

On peut remarquer que $\hat{\mathbb{V}}_{P|t}\left(\hat{t}_{y,\pi|t}\right)$ est également un estimateur sans biais de la variance non-conditionnelle de l'estimateur $\hat{t}_{y,\pi|t}$ si $\hat{t}_{y,\pi|t}$ est exactement sans biais. En effet, dans ce cas :

$$\begin{aligned}\mathbb{E}_P\left(\hat{\mathbb{V}}_{P|t}\left(\hat{t}_{y,\pi|t}\right)\right) &= \mathbb{E}_P\left(\mathbb{V}_{P|t}\left(\hat{t}_{y,\pi|t}\right)\right) && (2.45)\\ &= \mathbb{V}_P\left(\hat{t}_{y,\pi|t}\right). && (2.46)\end{aligned}$$

Lien avec les plans de sondage réjectifs
Les plans de sondage réjectifs sont des plans de sondage qui sont régis par une loi de probabilité conditionnelle. L'utilisation de l'estimateur par expansion conditionnel correspond à une inférence identique à celle d'un plan de sondage réjectif utilisant la

statistique ancillaire pour l'acceptation ou le rejet de l'échantillon. Tout se passe donc comme si notre échantillon provenait d'un plan de sondage réjectif où l'acceptation correspond à la condition (certes saugrenue)

$$\hat{t}_{x,\pi}(s) = t.$$

Les résultats développés pour le sondage réjectif (comme les estimateurs de variances où les estimateurs de probabilité d'inclusion) sont transférables à l'approche conditionnelle. On verra ainsi au chapitre 6 que des travaux sur le tirage de Poisson de taille fixe ou sur l'estimation de probabilités d'inclusion par des méthodes Monte Carlo (Fattorini, 2006) seront adaptés au contexte de l'inférence conditionnelle.

2.8.4 Les probabilités d'inclusion conditionnelles

L'estimateur par expansion conditionnel nécessite de connaître les probabilités d'inclusion conditionnelles. Pour un plan de sondage aléatoire simple, on sait faire un calcul exact pour un certain nombre de statistiques ancillaires (Rao, 1985) : taille d'un domaine, marges d'un tableau de contingence, tailles de strate.
Dans le chapitre 6 de cette thèse, on verra qu'on peut également calculer les probabilités d'inclusion conditionnelles pour le sondage de Poisson et le sondage de Poisson de taille fixe avec un conditionnement sur la taille de strates.

Tillé (1999) a proposé une méthode d'estimation des probabilités d'inclusion pour les échantillons de grande taille, reposant sur des hypothèses de normalité asymptotique des estimateurs par expansion.
Dans le chapitre 5, nous verrons une méthode alternative d'estimation des probabilités d'inclusion reposant sur des simulations Monte Carlo.

Bibliographie

Andersson, P. G. and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31(1) :95–99.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review / Revue Internationale de Statistique*, 51(3) :pp. 279–292.

Binder, D. A. (1991). Use of estimating functions for interval estimation from complex surveys. In *Proceedings of the survey research methods section*, pages 34–42. American Statistical Association.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples : a cookbook approach. *Survey Methodology*, 22 :17–22.

Binder, D. A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89(427) :1035–1043.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95(3) :555–571.

Demnati, A. and Rao, J. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30(1) :17–26.

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. In *Actes du colloque de la Société Statistique du Canada, Sherbrooke, Canada*.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, 25 :219–230.

Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des journées de méthodologie statistique*, pages 4–20.

Deville, J.-C., Sarndal, C., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, pages 1013–1020.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418) :376–382.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling : the cube method. *Biometrika*, 91(4) :893–912.

Dupont, F. (1996). Calage et redressement de la non-réponse totale. In *Actes des journées de méthodologie statistique, 15 et 16 décembre 1993*, number 56. INSEE-Méthodes.

Estevao, V. M. and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16(4) :379–400.

Estevao, V. M. and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2) :127–147.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs : A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2) :269–278.

Fuller, W. A. (2002). Regression estimation for survey samples (with discussion). *Survey Methodology*, 28(1) :5–23.

Fuller, W. A. (2009a). *Sampling Statistics*. Wiley.

Fuller, W. A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4) :933–944.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31 :1208–1211.

Hájek, J. (1971). Comment on" an essay on the logical foundations of survey sampling" by d. basu. *Foundations of Statistical Inference (eds VP Godambe and DA Sprott). Toronto : Holt, Rinehart, and Winston*.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *J. Am. Stat. Assoc.*, 78 :776–793.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, pages 663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377) :89–96.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1) :21–39.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2) :133.

Kott, P. S. (2009). Calibration weighting : Combining probability samples and linear prediction models. *Handbook of Statistics, Sample Surveys : Inference and Analysis*, 29B :55–82.

Krieger, A. M. and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, 18(2) :225–239.

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2005). Does the model matter ? comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7(3) :649–673.

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29 :33–44.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24 :51–55.

Lesage, E. and Haziza, D. (2013). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. *Submitted for publication*.

Little, R. J. A. (1986). Survey nonresponse adjustments. *International statistical review*, 54(1) :3.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Offcial Statistics*, 15(2) :305–327.

Montanari, G. E. (1987). Post-sampling efficient qr-prediction in large-sample surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 191–202.

Montanari, G. E. and Ranalli, M. G. (2002). Asymptotically efficient generalized regression estimators. *Journal of Official Statistics*, 18 :577–589.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā : The Indian Journal of Statistics, Series B*, pages 166–186.

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11 :15–31.

Robinson, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82(399) :826–831.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2) :377–387.

Royall, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, pages 657–664.

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76(373) :66–77.

Sautory, O. (2003). Calmar 2 : une nouvelle version du programme calmar de redressement d'échantillon par calage. In *Recueil : Symposium de Statistique Canada*.

Skinner, C. J., Holt, D., and Smith, T. F. (1989). *Analysis of complex surveys*. John Wiley & Sons.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2) :113–135.

Särndal, C.-E. and Lundström, S. (2010). Design for estimation : Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36 :131–144.

Särndal, C.-E., Swensson, B., and Wretman., J. (1992). *Model Assisted Survey Sampling*. New-Yok : Springer-Verlag.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities : Simple random sampling. *International Statistical Review*, 66 :303–322.

Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities : comlex design. *Survey Methodology*, 25(1) :57–66.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). Finite population sampling and inference : a prediction approach. *Recherche*, 67 :02.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, pages 411–414.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453) :185–193.

Chapter 3

Calibration on complex parameters

3.1 Introduction

The issue of complex parameters in calibrations has been discussed in the literature. Särndal (2007) reviewed a number of them, in particular the work of Harms and Duchesne (2006) on the calibration estimation of quantiles, and the work of Krapavickaite and Plikusas (2005) and Plikusas (2006) on calibration estimators of certain functions of totals. The originality of the approach in this article is that it reduces calibration on a complex parameter to calibration on a total for a new ad hoc auxiliary variable. The advantage of this approach is that current calibration tools can be used and that there is no need to solve a complex optimization program. In section 3.2 of the article, we review how the calibration method works, define calibration on complex parameters and describe simple cases in which calibration on a complex parameter can be reduced to calibration on a total. In section 3, we focus on parameters that can be defined as a solution to an estimating equation (Godambe and Thompson, 1986). We introduce the concept of calibration on a complex parameter defined by an estimating equation and show that the resulting calibration equation can be replaced with an equation for calibration on a total. Owen (1991(Owen,), 2001(Owen, 2001)) introduced the use of constraints based upon estimating equations of known parameters in the context of empirical likelihood.
In section 3.4, we present a general method of calibration on complex parameter that use linearisation techniques.

3.2 A Complex Parameter Defined as a Function of Totals

3.2.1 Review of Calibration on Totals

The calibration weights are obtained by solving the following optimization program:

$$\min_{\{w_k/\ k\in s\}} \sum_{k\in s} d(w_k, d_k)$$

under constraints:

$$\begin{cases} \hat{t}_{x1,w} = t_{x1} \\ \vdots \\ \hat{t}_{xP,w} = t_{xP} \end{cases}$$

$d(\cdot,\cdot)$ is a pseudo-distance, i.e., a function that measures the difference between the calibration weight and the sampling weight (unlike a difference, a pseudo-distance is not necessarily symmetrical on its two arguments). The program is solved with a Lagrangian. When the distance used is the χ^2 $\left(\text{i.e. } d(w_k, d_k) = \frac{1}{2}\frac{(w_k - d_k)^2}{d_k}\right)$ the solution is $w_k = d_k(1 + \boldsymbol{x}_k^T \boldsymbol{\lambda})$ (where $\boldsymbol{\lambda}$ is a P-vector of Lagrange multipliers).

3.2.2 Calibration on a Complex Parameter $\eta_{\mathbf{x}}$

Definition 1. *Let $x_1, ..., x_P$ be P auxiliary variables known on s, and let $\eta_{\mathbf{x}} = g(t_{x1}, ..., t_{xP})$ be a complex parameter, a function of the totals of those auxiliary variables, also known.*
In the case of calibration on the complex parameter $\eta_{\mathbf{x}}$, the calibration weights are obtained by solving the following optimization program:

$$\min_{\{w_k\}_{(k \in s)}} \sum_{k \in s} d(w_k, d_k)$$

under constraints:

$$\hat{\eta}_{\mathbf{x},CAL} = g(\hat{t}_{x1,CAL}, ..., \hat{t}_{xP,CAL}) = \eta_{\mathbf{x}}$$

The totals t_{xq} do not have to be known, but the complex parameter $\eta_{\mathbf{x}}$ does.

Consider the example of the ratio

$$R_{\mathbf{x}} = \frac{t_{x1}}{t_{x_2}} = \frac{\sum_{k \in U} x_{k,1}}{\sum_{k \in U} x_{k,2}}.$$

The calibration estimator of $R_{\mathbf{x}}$ is of the form

$$\widehat{R}_{\mathbf{x},CAL} = \frac{\sum_{k \in s} w_k x_{k,1}}{\sum_{k \in s} w_k x_{k,2}}.$$

The calibration equation in the case of calibration on a ratio is

$$\widehat{R}_{\mathbf{x},CAL} = \frac{\sum_{k \in s} w_k x_{k,1}}{\sum_{k \in s} w_k x_{k,2}} = R_{\mathbf{x}}.$$

$R_{\mathbf{x}}$ is known auxiliary information, as the total of the auxiliary variables usually is. This scenario may occur when we have proportions that are well known and stable over time, for example, but the specific totals in the numerator and denominator

are not known.
We described the case of calibration on a single complex parameter, but it is clearly a simple matter to calibrate on more than one complex parameter. In that case, there are as many constraints as calibration parameters.

3.2.3 Simple cases where calibration on a complex parameter can be reduced to calibration on a total

It is not easy to determine from the outset whether an equation for calibration on a complex parameter can be written in the form of an equation for calibration on a total. In other words, it is not always a trivial matter to find a "new" auxiliary variable z, associated with the complex parameter, on whose total we can calibrate.

For example, that is quite straightforward for all moments of an auxiliary variable x. If $\mu_{x^m} = \frac{1}{N}\sum_{k\in U} x_k^m$ is auxiliary information, we can simply take $z_k = x^m - \mu_{x^m}$ and calibrate on $t_z = 0 : \sum_{k\in s}(w_k x^m - \mu_{x^m}) = 0$.

If we want to calibrate on the variance and the mean of variable x with μ_x and σ_x^2 as auxiliary information, we can use the two new auxiliary variables

$$z_{k,1} = x_k - \mu_x$$

and

$$z_{k,2} = (x_k - \mu_x)^2 - \sigma_x^2.$$

On the other hand, if we do not know μ_x but we have σ_x^2 in the auxiliary information and we want to calibrate on that variance, things become more complicated. We can see this if we write the substitution estimator of σ_x^2 (where the sampling plan allows the population size N to be estimated exactly):

$$\hat{\sigma}^2_{x,CAL} = \frac{1}{N}\sum_{k\in s} w_k \left\{ x_k - \left(\frac{\sum_{l\in s} w_l x_l}{N} \right) \right\}^2 .$$

Finding a new auxiliary variable z is not straightforward, since the initial calibration equation is not linear relative to the weight vector. We will return to the variance case in section 3.3.3 below.

Ratio example

Proposition 2. *Calibration on a ratio is equivalent to calibration on the total of the new auxiliary variable:* $z_k = x_{k,1} - R_{\mathbf{x}} x_{k,2}$.
The calibration equation is written

$$\hat{t}_{z,CAL} = t_z = 0$$

Proof.

$$\begin{aligned}
\hat{t}_{z,CAL} = t_z &\iff \sum_{k\in s} w_k \left(x_{k,1} - R_{\mathbf{x}} x_{k,2}\right) = \sum_{k\in U} \left(x_{k,1} - R_{\mathbf{x}} x_{k,2}\right) \\
&\iff \hat{t}_{x1,CAL} - R_{\mathbf{x}} \hat{t}_{x2,CAL} = t_{x1,CAL} - R_{\mathbf{x}} t_{x2,CAL} = 0 \\
&\iff \frac{\hat{t}_{x1,CAL}}{\hat{t}_{x2,CAL}} = R_{\mathbf{x}}
\end{aligned}$$

i.e. $\widehat{R}_{\mathbf{x},CAL} = R_{\mathbf{x}}$. □

Function of a ratio of linear combinations of totals

Let $\eta_{\mathbf{x}}$ be a complex parameter that is a bijective function of a ratio of linear combinations of totals:

$$\eta_{\mathbf{x}} = h\left(\frac{\alpha' \cdot \mathbf{t_x}}{\beta' \cdot \mathbf{t_x}}\right) \tag{3.1}$$

where $\alpha' = (\alpha_1, ..., \alpha_P)$ and $\beta' = (\beta_1, ..., \beta_P)$ being vectors of real coefficients of size P, and $\mathbf{t_x}' = (t_{x^{(1)}}, ..., t_{x^{(P)}})$.

Proposition 3. *Performing a calibration on complex parameter $\eta_{\mathbf{x}}$ defined by function (3.1) is equivalent to calibrating on the total of the new auxiliary variable:*

$$z_k = \left\{\alpha' - h^{-1}\left(\eta_{\mathbf{x}}\right)\beta'\right\} \cdot \mathbf{x_k}$$

with calibration equation:

$$\hat{t}_{z,CAL} = \sum_{k\in s} w_k z_k = t_z = 0$$

Proof.

$$\begin{aligned}
\hat{\eta}_{\mathbf{x},CAL} = \eta_{\mathbf{x}} &\iff h\left(\frac{\alpha' \cdot \hat{\mathbf{t}}_{\mathbf{x},\mathbf{CAL}}}{\beta' \cdot \hat{\mathbf{t}}_{\mathbf{x},\mathbf{CAL}}}\right) = \eta_{\mathbf{x}} \\
&\iff \frac{\alpha' \cdot \hat{\mathbf{t}}_{\mathbf{x},\mathbf{CAL}}}{\beta' \cdot \hat{\mathbf{t}}_{\mathbf{x},\mathbf{CAL}}} = h^{-1}(\eta_{\mathbf{x}}) \\
&\iff \left(\alpha' - h^{-1}\left(\eta_{\mathbf{x}}\right)\beta'\right) \cdot \hat{\mathbf{t}}_{\mathbf{x},\mathbf{CAL}} = 0 \\
&\iff \sum_{k\in s} w_k \left(\alpha' - h^{-1}\left(\eta_{\mathbf{x}}\right)\beta'\right) \cdot \mathbf{x_k} = 0
\end{aligned}$$

□

Consider the example of the geometric mean:

$$\mu_{Geo,\mathbf{x}} = \left(\prod_{k\in U} x_k\right)^{1/N}$$

This expression can be rewritten as

$$\mu_{Geo,\mathbf{x}} = \exp\left(\frac{\sum_{k\in U}\ln(x_k)}{\sum_{k\in U}1}\right)$$

We denote $\mathbf{x}'_k = (x_{k,1}, x_{k,2}) = (\ln(x_k), 1)$, $\alpha' = (1,0)$, $\beta' = (0,1)$ and $h^{-1}(u) = \exp^{-1}(u) = \ln(u)$.

Hence, the new auxiliary variable is

$$z_k = \ln(x_k) - \ln(\mu_{Geo,\mathbf{x}}) \cdot 1$$

We will see later in the article that the estimating equations method provides another approach to displaying the new auxiliary variable(s) $\mathbf{z}$.

3.3 Parameter defined by an Estimating Equation

3.3.1 Estimating with an Estimating Equation

Certain parameters $\theta_{\mathbf{y}}$ are defined, or can be defined, as the solution to an implicit function known as the estimating equation on U (Godambe and Thompson 1986) de la forme :

$$\sum_{k\in U}\Phi(\theta_{\mathbf{y}}, \mathbf{y}_k) = 0$$

where $\mathbf{y}'_k = (y_{k,1}, ..., y_{k,Q})$ being the vector of values taken by the variables of interest for individual k.
In this context, an estimator of $\theta_{\mathbf{y}}$ is defined for sample s, denoted $\hat{\theta}_{\mathbf{y},ee,\pi}$ which is the solution of the estimating equation on s (see in particular (Hidiroglou et al., 2002)):

$$\sum_{k\in s} d_k\Phi(\hat{\theta}_{\mathbf{y},ee,\pi}, \mathbf{y}_k) = 0.$$

Table 3.1: Examples of Parameters defined by Estimating Equations on U

Parameter	$\Phi(\theta_{\mathbf{y}}, \mathbf{y}_k)$	Estimating Equation on U
moyenne μ	$(y_k - \mu)$	$\sum_{k\in U}(y_k - \mu) = 0$
ratio $R = \frac{\mu_1}{\mu_2}$	$(y_k^{(1)} - Ry_k^{(2)})$	$\sum_{k\in U}(y_k^{(1)} - Ry_k^{(2)}) = 0$
médiane m	$(\mathbb{1}_{y_k\leq m} - \frac{1}{2})$	$\sum_{k\in U}(\mathbb{1}_{y_k\leq m} - \frac{1}{2}) = 0$

Consider also the example of the coefficient of a logistic regression. Let y_1 be a dichotomous variable that takes the values 0 and 1 on U, and let y_2 be a quantitative

variable. The value $Y_{k,1}$ taken by $y^{(1)}$ for unit k is assumed to be an instance of the random variable $Y_k^{(1)}$, which has a Bernoulli distribution

$$\mathfrak{B}\left(1, p_k = \frac{1}{1+\exp\left(-\beta_0 Y_{k,2}\right)}\right).$$

We have limited the number of parameters to one, but it would be just as simple to consider the multidimensional case. However, we should provide a definition of the estimating equations that take the case of the vector parameters into account.
The parameter of interest to us is the estimator of β_0, denoted β, calculated on the finite population by the maximum likelihood method. The estimating equation of β on U will be the maximum likelihood equation. The loglikelihood in the case of Bernoulli variables is

$$\mathbf{L}(\beta) = \sum_{k\in U} Y_{k,1} \ln\left(p_k\right) + \sum_{k\in\mathcal{U}} \left(1 - Y_{k,1}\right) \ln\left(1 - p_k\right).$$

It is easy to derive the estimating equation of β on U:

$$\sum_{k\in\mathcal{U}} Y_{k,2}\left(Y_{k,1} - \frac{1}{1+\exp\left(-\beta Y_{k,2}\right)}\right) = 0$$

The estimating equation on s which defines the estimator$\hat{\beta}_{ee,\pi}$ on the basis of the sampling weights is

$$\sum_{k\in s} d_k Y_{k,2}\left(Y_{k,1} - \frac{1}{1+\exp\left(-\hat{\beta}_{ee,\pi} Y_{k,2}\right)}\right) = 0.$$

The estimating equation is not linear in the parameter; $\hat{\beta}_{ee,\pi}$ cannot be expressed as a simple function of the observations.
The logistic regression example is very interesting because it shows that we do not need to know $\hat{\beta}_{ee,\pi}$ to perform the calibration. We will see in the next subsection that we only need to know the generic term of the estimating equation on $\mathcal{U}$

$$\Phi(\beta, \mathbf{y}_k) = Y_{k,2}\left(Y_{k,1} - \frac{1}{1+\exp\left(-\beta Y_{k,2}\right)}\right),$$

for all $k \in s$.

3.3.2 Calibration in the Case of Parameters defined by Estimating Equations

Let $\mathbf{x}_k' = (x_1, ..., x_P)$ be the vector of P known auxiliary variables on s, and let $\eta_{\mathbf{x}}$ be a complex parameter, also known, defined by the estimating equation

$$\sum_{k\in\mathcal{U}} \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0.$$

Definition 2. *In the case of calibration on the complex parameter $\eta_{\mathbf{x}}$ the calibration weights are obtained by solving the following optimization program:*

$$\min_{\{w_k,\, k\in s)\}} \sum_{k\in s} d(w_k, d_k)$$

under constraints: $\sum_{k\in s} w_k \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0$

Proposition 4. *Calibration on a complex parameter $\eta_{\mathbf{x}}$ defined by an estimating equation, is equivalent to a calibration on the total of the new auxiliary variable: $z_k = \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k)$ with the calibration constraint $\sum_{k\in s} w_k z_k = 0$.*

Definition 3. *A calibration estimator of the parameter of interest $\theta_{\mathbf{y}}$ denoted ,$\hat{\theta}_{\mathbf{y},ee,CAL}$ is a solution to the estimating equation on s weighted by the calibration weights $\{w_k\}_{(k\in s)}$*

$$\sum_{k\in s} w_k \Phi(\hat{\theta}_{\mathbf{y},ee,CAL}, \mathbf{y}_k) = 0.$$

In most cases, the solution to the estimating equation is unique. The median is an example of a parameter for which there may be more than one solution. In this case, the infimum is often used as an estimator.

Proposition 5. *If there is only one solution to the equation*

$$\sum_{k\in s} w_k \Psi(\hat{\eta}_{\mathbf{x},ee,CAL}, \mathbf{x}_k) = 0,$$

then

$$\hat{\eta}_{\mathbf{x},ee,CAL} = \eta_{\mathbf{x}}.$$

Proof. $\eta_{\mathbf{x}}$ is a solution to the estimating equation that defines $\hat{\eta}_{\mathbf{x},ee,CAL}$. Since there is a unique solution, we have $\hat{\eta}_{\mathbf{x},ee,CAL} = \eta_{\mathbf{x}}$. □

3.3.3 Calibration on a variance

In this section, we examine calibration on variance σ_x^2 which is a more complicated complex parameter than those discussed above. We will show that when the variance is the only auxiliary information we have, we can perform an approximate calibration that produces calibration weights that have better properties than the sampling weights.

Back to the variance case. The mean μ_x and the variance σ_x^2 on $\mathcal{U}$ of auxiliary variable x can be defined by two estimating equations on $\mathcal{U}$:

$$\begin{cases} \sum_{k\in U} (x_k - \mu_x) = 0 & (3.2) \\ \sum_{k\in U} \left\{ (x_k - \mu_x)^2 - \sigma_x^2 \right\} = 0 & (3.3) \end{cases}$$

If we know the two parameters, calibrating on them is easy, since we merely have to calibrate on the totals of the two new auxiliary variables $z^{(1)} = x - \mu_x$ et $z^{(2)} = (x - \mu_x)^2 - \sigma_x^2$.

On the other hand, if we consider the textbook case where the mean μ_x is not known, the parameter σ_x^2 cannot be defined by a unique estimating equation. If we replace μ_x with its explicit definition

$$\mu_x = \frac{\sum_{l\in U} x_l}{\sum_{j\in U} 1}$$

in equation (3.3), we obtain the equation

$$\sum_{k\in U} \left\{ \left(x_k - \frac{\sum_{l\in U} x_l}{\sum_{j\in U} 1} \right)^2 - \sigma_x^2 \right\} = 0$$

which cannot be written in the form of an estimating equation:
$\sum_{k\in U} \Psi(\sigma_x^2, x_k) = 0$.

μ_x thus becomes a nuisance parameter (Binder, 1991). To overcome this difficulty, we can replace it in equation (3.3) with its substitution estimator: $\hat{\mu}_{x,\pi} = \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}$ where $\hat{N}_\pi = \sum_{k\in s} d_k 1$ being the Horvitz-Thompson estimator of the size of population $\mathcal{U}$. This leads to the "approximate" calibration equation

$$\sum_{k\in s} w_k \left\{ \left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right)^2 - \sigma_x^2 \right\} = 0. \quad (3.4)$$

Proposition 6. *With estimating equation (3.4), calibration on the variance is not perfect, and we have*

$$\hat{\sigma}^2_{x,ee,CAL} = \sigma_x^2 - \left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right)^2 \quad (3.5)$$

Proof. – The "approximate" calibration equation is equation(3.4).
– The definition of the parameters' calibration estimators:

$$\begin{cases} \sum_{k\in s} w_k \left(x_k - \hat{\mu}_{x,ee,CAL} \right) = 0 \\ \sum_{k\in s} w_k \left(\left(x_k - \hat{\mu}_{x,ee,CAL} \right)^2 - \hat{\sigma}^2_{x,ee,CAL} \right) = 0 \end{cases}$$

This can be rewritten

$$\begin{cases} \hat{\mu}_{x,ee,CAL} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k} = \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \\ \sum_{k\in s} w_k \left(\left(x_k - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right)^2 - \hat{\sigma}^2_{x,ee,CAL} \right) = 0 \end{cases}$$

– If we subtract the second estimating equation from the approximate calibration equation, we get

$$\sum_{k\in s} w_k \left(\left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}} \right)^2 - \left(x_k - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right)^2 - \sigma^2_x + \hat{\sigma}^2_{x,ee,CAL} \right) = 0$$

Using the identity $a^2 - b^2 = (a-b)(a+b)$, we have

$$\sum_{k\in s} w_k \left(\left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right) \left(2x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right) \right) - \hat{N}_{CAL} \left(\sigma^2_x - \hat{\sigma}^2_{x,ee,CAL} \right) = 0$$

$$\left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right) \sum_{k\in s} w_k \left(2x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right) - \hat{N}_{CAL} \left(\sigma^2_x - \hat{\sigma}^2_{x,ee,CAL} \right) = 0$$

$$\left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right) \left(2\hat{t}_{x,CAL} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \hat{N}_{CAL} - \hat{t}_{x,CAL} \right) - \hat{N}_{CAL} \left(\sigma^2_x - \hat{\sigma}^2_{x,ee,CAL} \right) = 0$$

$$\hat{N}_{CAL} \left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right)^2 - \hat{N}_{CAL} \left(\sigma^2_x - \hat{\sigma}^2_{x,ee,CAL} \right) = 0$$

This is the same as the expression for $\hat{\sigma}^2_{x,ee,CAL}$ in equation (3.5).

□

This result is interesting because, without an exact calibration, we have a calibration estimator of σ^2_x that is asymptotically more precise than the substitution estimator $\hat{\sigma}^2_{x,\pi}$. That is, if we resort to the asymptotic framework typically used in surveys and employ linearization of complex estimators Deville(1999), we have for $\hat{\sigma}^2_{x,\pi}$:

$$\hat{\sigma}^2_{x,\pi} - \sigma^2_x = O_p\left(\frac{1}{\sqrt{n}} \right)$$

and for $\hat{\sigma}^2_{x,ee,CAL}$:

$$\left(\hat{\sigma}^2_{x,ee,CAL} - \sigma^2_x \right)^{1/2} = \left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right) = O_p\left(\frac{1}{\sqrt{n}} \right)$$

This yields

$$\hat{\sigma}^2_{x,ee,CAL} - \sigma^2_x = O_p\left(\frac{1}{n} \right)$$

Approximate calibration estimation yields a gain in probability order of magnitude.

We can mention another interesting approach in order to calibrate on variance. We still use the definition of the parameters' calibration estimators:

$$\begin{cases} \hat{\mu}_{x,ee,CAL} = \frac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k} \\ \sum_{k\in s} w_k \left(x_k - \hat{\mu}_{x,ee,CAL}\right)^2 = \left(\sum_{k\in s} w_k\right) \sigma_x^2. \end{cases}$$

We then obtain a non-linear calibration equation:

$$\left(\sum_{k\in s} w_k x_k^2\right)\left(\sum_{k\in s} w_k\right) - \left(\sum_{k\in s} w_k x_k\right)^2 = \sigma_x^2 \left(\sum_{k\in s} w_k\right)^2.$$

Using a functional approach for calibration, we look for a set of calibration weights of the form

$$w_k = F(\lambda x_k),$$

where $F(.)$ is one of the usual calibration function. This yields the equation

$$\begin{aligned} \Phi(\lambda) &= \left(\sum_{k\in s} F(\lambda x_k) x_k^2\right)\left(\sum_{k\in s} F(\lambda x_k)\right) \\ &- \left(\sum_{k\in s} F(\lambda x_k) x_k\right)^2 - \sigma_x^2 \left(\sum_{k\in s} F(\lambda x_k)\right)^2 \\ &= 0, \end{aligned}$$

that we can try to solve with a Newton-Raphson method. This idea could lead to other research investigations.

3.4 Any parameters: linearization approach

In this section we provide the statistician with a generalized method to calibrate on complex parameters. With this approach, we will show that the calibration is not always exact.

Consider η_x, a parameter of degree α, and z_k his linearised variable. We know from (2.14) that:

$$\frac{\sqrt{n}}{N^\alpha}\left(\hat{\eta}_{x,CAL} - \eta_x\right) = \frac{\sqrt{n}}{N^\alpha}\left(\hat{t}_{z,CAL} - t_z\right) + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{3.6}$$

The linearization approach for calibration on complex parmeter consists of calibrate on the total of the linearized variable of η. Calibration equation used to calculate the calibration weights w_k, $k \in s$ are

$$\sum_{k\in s} w_k z_k = t_z. \tag{3.7}$$

It follows from (3.6) that $\hat{\eta}_{x,CAL}$, the calibration estimator of η_x, verifies

$$N^{-\alpha}\left(\hat{\eta}_{x,CAL}-\eta_x\right)=O_p\left(\frac{1}{n}\right), \tag{3.8}$$

which means that $\hat{\eta}_{x,CAL}$ is n-consistant whereas expansion estimator $\hat{\eta}_{x,\pi}$ is $\sqrt{n}$-consistant.

3.4.1 linearized calibration for the variance parameter σ_x^2

The variance of the variable x is a function of totals and can be defined explicitely by:

$$\sigma_x^2=\frac{1}{N}\sum_{k\in U}x_k^2-\left(\frac{\sum_{k\in U}x_k}{N}\right)^2=f(t_x,t_{x^2},N).$$

From (2.13), it follows that the linearized variable of σ_x^2 is:

$$z_k=-2\frac{t_x}{N^2}x_k+\frac{1}{N}x_k^2-\frac{1}{N}\left\{\sigma_x^2-\left(\frac{t_x}{N}\right)^2\right\}. \tag{3.9}$$

When parameters N and t_x are unknown, we can use estimations of these parmeters and take an approximated linearized variable denoted $\tilde{z}_k$:

$$\tilde{z}_k=-2\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi^2}x_k+\frac{1}{N}x_k^2-\frac{1}{\hat{N}_\pi}\left\{\sigma_x^2-\left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)^2\right\}. \tag{3.10}$$

If we calibrate on the total of the linearized variable (3.9) of σ_x^2, we have

$$\begin{aligned}\hat{\sigma}_{x,CAL}^2 &= \sigma_x^2\\ &+ \frac{N}{N-1}\left\{\frac{\hat{N}_{CAL}-N}{N}\left(\frac{\hat{t}_{x^2,CAL}}{\hat{N}_{CAL}}-\frac{t_{x^2}}{N}\right)+\frac{\hat{N}_{CAL}}{N}\left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}}-\frac{t_x}{N}\right)^2\right\},\end{aligned}$$

which confirm that the linearized calibration is not always exact and that the order of probability of the différence between $\hat{\sigma}_{x,CAL}^2$ and σ_x^2 is $O_p\left(\frac{1}{n}\right)$.

3.4.2 Gini index example

Many definitions of the Gini index exists (Berger, 2008). Let's take the definition from Glasser (1962):

$$G=\frac{1}{t_x}\sum_{k\in\mathcal{U}}\left(2F(x_k)-1\right)x_k, \tag{3.11}$$

where $F(x)$ is the cumulative density function of the variable x defined by

$$F(x)=\frac{1}{N}\sum_{l\in\mathcal{U}}\mathbb{1}_{x_l\leq x}. \tag{3.12}$$

A plug-in estimator of G is given by Kovacevic and Binder (1997):

$$\hat{G}_\pi = \frac{1}{\hat{t}_{x,\pi}} \sum_{k \in s} d_k \left(2\hat{F}_\pi(x_k) - 1\right) x_k, \tag{3.13}$$

where

$$\hat{F}_\pi(x) = \frac{1}{\hat{N}_\pi} \sum_{l \in s} d_l \mathbb{1}_{x_l \leq x}. \tag{3.14}$$

Deville (1996 et 1999) gave an expression of the linearized variable for the Gini index:

$$z_k = \frac{1}{t_x} \left[\left(2F(x_k) - (G+1)\right) x_k + \frac{2}{N} \sum_{l \in U} \mathbb{1}_{x_k \leq x_l} x_l - (G+1)\frac{t_x}{N} \right]. \tag{3.15}$$

If auxiliary information provides the value of G but not the cumulative density function $F(x)$ and the value of $\sum_{l \in U} \mathbb{1}_{x_k \leq x_l} x_l$, then it is possible to use the following approximated linearized variable of (3.15):

$$\tilde{z}_k = \frac{1}{\hat{t}_{x,\pi}} \left[\left\{2\hat{F}_\pi(x_k) - (G+1)\right\} x_k + \frac{2}{\hat{N}_\pi} \sum_{l \in s} d_k \mathbb{1}_{x_k \leq x_l} x_l - (G+1)\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right]. \tag{3.16}$$

3.5 Conclusion

In this article, we presented a simple method of performing a calibration in cases where the auxiliary information takes the form of a complex parameter. That method is based on the concept of the estimating equation. Its major advantage is that it can be used with current calibration software.
In future research, it would be interesting to determine the practical cases in which the use of complex parameters in the calibration improves the precision of the parameters of interest.

At last, it is interesting to mention that all the results presented in this article for calibration can be transposed to balanced sampling. Balanced equations and calibration equations are playing an equivalent role; the former at the sampling stage and the latter at the estimation stage.

Bibliography

Berger, Y. G. (2008). A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient.

Binder, D. A. (1991). Use of estimating functions for interval estimation from complex surveys. In *Proceedings of the survey research methods section*, pages 34–42. American Statistical Association.

Deville, J.-C. (1996). Estimation de la variance du coefficient de Gini estimé par sondage. *Actes des journées de méthodologie statistique*, pages 269–288.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:219–230.

Glasser, G. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, pages 648–654.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54:127–138.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32:37–52.

Hidiroglou, M. H., Rao, J. N. K., and Yung, W. (2002). Estimating equations for the analysis of survey data using poststratification information. *Survey Methodology*, 64(2):364–378.

Kovacevic, M. and Binder, D. (1997). Variance estimation for measures of income inequality and polarization-the estimating equations approach. *Journal of Official Statistics*, 13(1):41–58.

Krapavickaite, D. and Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica*, 16:347–364.

Owen, A. B. Empirical likelihood for linear models. *The Annals of Statistics*, (19):1725–1747.

Owen, A. B. (2001). *Empirical Likelihood*. New York: Chapman and Hall.

Plikusas, A. (2006). Non-linear calibration. In *Proceedings, Workshop on survey sampling*, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):113–135.

Chapter 4

A discussion of weighting procedures for unit nonresponse

4.1 Introduction

Weighting procedures are commonly applied in surveys to compensate for nonsampling errors such as nonresponse errors and coverage errors. In a recent paper, Brick (2013) provided an excellent overview of weighting in the presence of unit nonresponse; see also Kalton and Flores-Cervantes (2003). In this note, we highlight some aspects of two types of weight adjustment procedures commonly used in practice in the context of unit nonresponse: (i) nonresponse propensity weighting followed by calibration and (ii) nonresponse calibration weighting.

In the case of nonresponse propensity weighting followed by calibration, also known as the two-step approach, the weights are adjusted in two distinct steps: the basic (design) weights of respondents are first multiplied by a nonresponse adjustment factor, which is defined as the inverse of the estimated response probability. The adjusted weights are further modified so that survey weighted estimates agree with known population totals. In the first step, survey statisticians aim at reducing the nonresponse bias, which may be appreciable when respondents and nonrespondents are different with respect to the survey variables. Whether or not one will succeed in achieving an efficient bias reduction depends on the availability of powerful auxiliary information, which is a set of variables available for both respondents and nonrespondents. In the second step, some form of calibration (e.g., post-stratification) is performed in order to ensure consistency between survey weighted estimates and known population totals. Calibration procedures require auxiliary variables (called calibration variables) available for the respondents and such that their population totals are known. In practice, the calibration variables are often specified by survey managers, who wish to ensure consistency with respect to some important variables (e.g., age and sex). Moreover, if the calibration variables are related to the characteristics of interest, the resulting calibration estimators tend to be more efficient than the non calibrated ones.

Nonresponse calibration weighting, also known as the single-step approach, uses cal-

ibration with three simultaneous goals in mind: reduce the nonresponse bias, ensure consistency between survey estimates and known population totals and, possibly, reduce the variance of point estimators. Unlike in the two-step approach, explicit estimation of the response probabilities is not required. In the absence of nonsampling errors, calibration consists of determining a set of calibrated (or final) weights as close as possible to the basic weights, while satisfying calibration constraints. A calibrated weight is expressed as the basic weight multiplied by a calibration adjustment factor, which depends on a calibration function. Commonly used calibration functions include the linear function, the exponential function, the truncated linear function and the logit function. Deville and Särndal (1992) showed that these distance functions are asymptotically equivalent in the sense that they all lead to the calibration estimator based on the linear calibration function. In the absence of nonsampling errors, calibration estimators are asymptotically unbiased and the calibration function is usually chosen so that the distribution of calibrated weights is "cosmetically attractive". For example, a problem that can be encountered with the linear function is the occurrence of negative weights, which can be prevented by using the exponential function that ensures positive weights. However, the latter may lead to extreme weights, which in turns may contribute to increase the instability of point estimators for characteristics of interest weakly correlated with the calibration variables. In this case, functions such as the truncated linear function or the logit function can be used in order to ensure that the calibration adjustment factors lie between pre-specified lower and upper bounds. In the presence of unit nonresponse, the situation is more subtle as different calibration functions may lead to calibration estimators with substantially different properties in terms of bias and variance. While the choice of calibration variables has been widely discussed in the literature (e.g., Särndal and Lundström, 2005 and Särndal, 2011), how to select an appropriate calibration function in the presence of unit nonresponse and the effect of function misspecification have not been fully discussed. Two notable exceptions are Kott (2006) and Kott and Liao (2012).

In this note, we argue that, even though nonresponse calibration weighting does not explicitly use estimated response probabilities in the construction of point estimators, a complete modeling exercise of these probabilities is unavoidable in order to ensure that an appropriate calibration function is selected. Failing to do so may lead to substantially biased calibration estimators (sometimes exhibiting a bias larger than that of unadjusted estimators) as we show empirically in Section 4.

4.2 Nonresponse propensity weighting followed by calibration

Let $U = \{1, 2, ..., N\}$ be a finite population consisting of N elements. We consider the problem of estimating a population total $t_y = \sum_{k \in U} y_k$, where y_k denotes the k-th value of the characteristic of interest y, $k = 1, ..., N$. A sample s, of size n, is selected from U according to a given sampling design $p(s)$. Let π_k denote the first-order inclusion probability of unit k in the sample and $d_k = 1/\pi_k$ denote its

design weight. In the presence of unit nonresponse, the characteristics of interest are observed for a subset, s_r, of the original sample s. Let ϕ_k be the unknown response propensity attached to unit k. We assume that $\phi_k > 0$ for all k and that units respond independently of one another. Also, we postulate the following nonresponse model

$$\phi_k = m(\mathbf{z}_k, \boldsymbol{\gamma}) \tag{4.1}$$

for some function $m(.)$, where $\mathbf{z}_k$ is a vector of auxiliary variables available for both respondents and nonrespondents and $\boldsymbol{\gamma}$ is a vector of unknown parameters. Ideally, the $\mathbf{z}$-vector should include variables that are related to both the response propensity and the characteristics of interest. If an auxiliary variable z is related to the response propensity but is unrelated to the characteristics of interest, it should be excluded from (4.1) since it will not help reducing the nonresponse bias and is likely to contribute in increasing the dispersion of the adjusted weights, which in turns may lead to potentially inefficient estimators (e.g., Little and Vartivarian, 2005).

An estimate of ϕ_k is $\hat{\phi}_k = m(\mathbf{z}_k, \hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is an estimator of $\boldsymbol{\gamma}$ (for example, the maximum likelihood estimator). The adjusted weights for nonresponse are defined as $w_k^* = d_k/\hat{\phi}_k$ for $k \in s_r$. Applying these weights to a characteristic of interest y leads to the Propensity Score Adjusted (PSA) estimator of t_y:

$$\hat{t}_{PSA} = \sum_{k \in s_r} d_k \hat{\phi}_k^{-1} y_k = \sum_{k \in s_r} w_k^* y_k. \tag{4.2}$$

The rationale behind this type of weighting procedure is similar in spirit to weighting for two-phase sampling. The PSA estimator (4.2) is asymptotically unbiased and consistent for t_y *regardless of the characteristic* y *being estimated* if (4.1) is correctly specified, which entails selecting the appropriate vector of auxiliary variables $\mathbf{z}$ as well as correctly specifying the form of $m(.)$. Examples of parametric nonresponse models include logit and probit models; see Kim and Kim (2007) for the theoretical properties of the PSA estimator in the case of parametric nonresponse models. PSA estimation based on parametric models is rarely used in statistical agencies because the resulting estimators are vulnerable to misspecification of the form of $m(.)$; e.g., Da Silva and Opsomer (2006). A popular method in practice consists of first obtaining estimated response probabilities $\hat{\phi}_k$ using a parametric model (e.g., the logit model) and partitioning the sample into homogeneous weighting classes formed on the basis of these estimated response probabilities. The basic weight of a respondent in a given class is then adjusted using the observed response rate within the same class; e.g., Little (1986), Eltinge and Yansaneh (1997) and Haziza and Beaumont (2007). This method is nonparametric in nature and is expected to provide a certain degree of robustness if the form of $m(.)$ is misspecified. Other nonparametric methods include smoothing methods such as kernel and local polynomial methods (e.g., Giommi, 1987 and Da Silva and Opsomer, 2006, 2009) and regression trees (e.g., Phipps and Toth, 2012).

The adjusted weights w_k^* are further modified so that survey weighted estimates agree with known population totals. More specifically, we assume that a vector of

calibration variables $\mathbf{x}^*$ is available for $k \in s_r$ and that the vector of population totals $\mathbf{t}_{\mathbf{x}^*} = \sum_{k \in U} \mathbf{x}_k^*$ is known. The final weight attached to unit k is defined as

$$w_k = w_k^* F(\hat{\boldsymbol{\lambda}}_*^\top \mathbf{x}_k^*), \tag{4.3}$$

where $F(.)$ is a monotonic and twice differentiable function such that $F(0) = 1$ and $F'(0) = 1$. The final weights w_k satisfy the calibration constraints

$$\sum_{k \in s_r} w_k \mathbf{x}_k^* = \mathbf{t}_{\mathbf{x}^*}. \tag{4.4}$$

The weight w_k in (4.3) is the product of the adjusted weight w_k^* and the calibration adjustment factor $F(\hat{\boldsymbol{\lambda}}_*^\top \mathbf{x}_k^*)$. Linear weighting is a special case of (4.3) for which the weights w_k are given by

$$w_k = w_k^*(1 + \hat{\boldsymbol{\lambda}}_*^\top \mathbf{x}_k^*).$$

Another popular weighting method is exponential weighting for which the weights w_k are given by

$$w_k = w_k^* \exp\left(\hat{\boldsymbol{\lambda}}_*^\top \mathbf{x}_k^*\right).$$

Alternative weighting methods are discussed in Deville and Särndal (1992) and Kott and Liao (2012), among others. Applying the final weights w_k to a characteristic of interest y leads to the two-step calibration estimator

$$\hat{t}_{C,2} = \sum_{k \in s_r} w_k y_k. \tag{4.5}$$

If the nonresponse model (4.1) is correctly specified, then $\hat{t}_{C,2}$ is asymptotically unbiased for t_y regardless of the characteristic y being estimated. Moreover, if the $\mathbf{x}^*$-vector is linearly related to y, then $\hat{t}_{C,2}$ is expected to be more efficient than $\hat{t}_{PSA}$.

4.3 Nonresponse calibration weighting

Following Särndal and Lundström (2005), we distinguish between two levels of auxiliary information:

(1) U-level: a vector of auxiliary variables $\mathbf{x}_k^*$ for $k \in s_r$ and the vector of population totals $\mathbf{t}_{\mathbf{x}^*} = \sum_{k \in U} \mathbf{x}_k^*$ is known.

(2) s-level: a vector of auxiliary variables $\mathbf{x}_k^o$ is available for $k \in s$ but the vector of population totals, $\sum_{k \in U} \mathbf{x}_k^o$, is unknown. Instead, the vector of complete data estimators, $\hat{\mathbf{t}}_{\mathbf{x}^o} = \sum_{k \in s} d_k \mathbf{x}_k^o$, is available.

We define the stacked vector of auxiliary variables for unit k as $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}$. The final weights $\tilde{w}_k$ are given by

$$\tilde{w}_k = d_k F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k), \tag{4.6}$$

where $\hat{\boldsymbol{\lambda}}_r$ is determined so that the calibration constraints

$$\sum_{k \in s_r} \tilde{w}_k \mathbf{x}_k = \mathbf{t}_{\mathbf{x}}$$

are satisfied, with $\mathbf{t}_{\mathbf{x}} = (\mathbf{t}_{\mathbf{x}^*}, \hat{\mathbf{t}}_{\mathbf{x}^o})^\top$. The final weight $\tilde{w}_k$ in (4.6) is the product of the design weight d_k and the nonresponse/calibration adjustment factor $F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k)$. Applying the weights $\tilde{w}_k$ to a characteristic of interest y leads to the one-step calibration estimator

$$\hat{t}_{C,1} = \sum_{k \in s_r} \tilde{w}_k y_k = \sum_{k \in s_r} d_k F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k) y_k. \tag{4.7}$$

Suppose that only U-level information is available. Then, regardless of the choice of the calibration function $F(.)$, the one-step calibration estimator $\hat{t}_{C,1}$ perfectly estimates the true population total t_y if $y_k = \mathbf{x}_k^\top \boldsymbol{\beta}$ for some vector $\boldsymbol{\beta}$. Hence, we expect $\hat{t}_{C,1}$ to exhibit a small bias if the characteristic of interest y and the $\mathbf{x}$-vector are linearly related and the relationship is strong.

In multipurpose surveys, the number of variables of interest is typically large (possibly few hundred) and many variables collected are categorical rather than continuous. Therefore, in most situations encountered in practice, it is unrealistic to presume that the $\mathbf{x}$-vector is linearly related to all y-variables, in which case some estimates could suffer from bias. Recall that the PSA estimator (4.2) is asymptotically unbiased for t_y regardless of the characteristic y being estimated. A comparison of (4.2) and (4.7) suggests that $\hat{t}_{C,1}$ is asymptotically unbiased for t_y for any characteristic of interest y if $F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k)$ is a good estimate ϕ_k^{-1}. In other words, the nonresponse/calibration adjustment factor $F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k)$ can be viewed as an estimate of ϕ_k^{-1}.

Using a first-order Taylor expansion and ignoring higher order terms, the asymptotic conditional nonresponse bias of $\hat{t}_{C,1}$ is given by

$$Bias(\hat{t}_{C,1}) \;=\; -\sum_{k \in U} (1 - \phi_k F_k) \left(y_k - \mathbf{x}_k^\top \boldsymbol{B}_{p,f}\right), \tag{4.8}$$

where

$$\boldsymbol{B}_{pf} = \left(\sum_{k \in U} \phi_k f_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \sum_{k \in U} \phi_k f_k \mathbf{x}_k y_k$$

with $F_k \equiv F\left(\boldsymbol{\lambda}_N^\top \mathbf{x}_k\right)$ and $f_k \equiv F'\left(\boldsymbol{\lambda}_N^\top \mathbf{x}_k\right)$ with $\boldsymbol{\lambda}_N$ denoting the probability limit of $\hat{\boldsymbol{\lambda}}_r$. Expression (4.8) suggests that the asymptotic bias vanishes if the residuals $e_k = \left(y_k - \mathbf{x}_k^\top \boldsymbol{B}_{p,f}\right)$ are unrelated to $\phi_k F_k$. This condition is satisfied if

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \epsilon_k \tag{4.9}$$

with

$$\mathbb{E}(\epsilon_k | \mathbf{x}_k) = 0. \tag{4.10}$$

Alternatively, it is satisfied if

$$F_k = \phi_k^{-1}. \tag{4.11}$$

In multipurpose surveys, it is unrealistic to presume that model (4.9) and/or (4.11) hold for every characteristic of interest y. In this case, some estimates will suffer from potential bias. In contrast, selecting a calibration function $F(.)$ such that (4.11) is satisfied, ensures that the one-step calibration estimator is asymptotically unbiased regardless of the characteristic of interest y.

For linear weighting, it follows from (4.11) that $\hat{t}_{C,1}$ is asymptotically unbiased for t_y for every y if

$$\phi_k^{-1} = 1 + \boldsymbol{\lambda}^\top \mathbf{x}_k \quad \text{for all } k \in U, \tag{4.12}$$

for a vector of unknown constants $\boldsymbol{\lambda}$; see Särndal and Lundström (2005, Chapter 9). For exponential weighting, we require

$$\phi_k^{-1} = \exp\left(\lambda^\top \mathbf{x}_k\right) \quad \text{for all } k \in U; \tag{4.13}$$

see also Kott and Liao (2012) for a discussion alternative weighting methods. This begs the following questions: if we use linear weighting (respectively exponential weighting), how make sure that the form (4.12) (respectively (4.13)) is reasonable? How is $\hat{t}_{C,1}$ affected if (4.12) (respectively (4.13)) is not an appropriate description of the relationship linking the $\mathbf{x}$-vector and the ϕ_k's?

A key aspect here is to realize that each calibration function corresponds to a specific parametric nonresponse model. By choosing a given calibration function, one is effectively making a strong statement about the underlying nonresponse mechanism. Therefore, in order to avoid an incorrect functional form, it is necessary to perform a complete modeling exercise in order to validate the form of the function linking the response probability ϕ_k to the vector of auxiliary variable $\mathbf{x}_k$. Unlike in the complete data situation, the choice of the calibration function must be based on statistical considerations rather than "cosmetic" considerations. Failing to do so may result in highly biased estimators as shown empirically in the next section. Also, there may not exist a calibration function that corresponds to the inverse of the estimated response probabilities. For instance, suppose that the relationship between the response probability and a single auxiliary variable x is described by a non-monotone function (see Example 3 in Section 4). In this case, it may be difficult to find a calibration function which provides an adequate description of the inverse of the response probability, which in turns may translate into bias.

4.4 Simulated examples

We conducted a simulation study to illustrate the importance of carefully selecting a calibration function $F(\cdot)$. We generated three populations of size $N = 1\ 000$, each consisting of a variable of interest y and an auxiliary variable x. In all three populations, the x-values were first generated from a uniform distribution $(0, 80)$. The model used to generate the y-values are given in Example 1-Example 3 below.

In order to focus on the nonresponse error, we considered the census case; i.e., $n = N = 1\ 000$ and $d_k = 1$ for all k. In each population, units were assigned a response probability ϕ_k according to a given nonresponse mechanism. For each mechanism, the parameters were set so that the overall response rate was approximately equal to 50%. The nonresponse mechanisms used in each example are presented in Table 4.1.

The response indicators R_k for $k \in U$ were generated independently from a Bernoulli distribution with parameter ϕ_k, resulting in a population of respondents U_r of size N_r. The nonresponse process was repeated $J = 1\ 000$ times, leading to $J = 1\ 000$ sets of respondents for each nonresponse mechanism.

To estimate t_y, we computed two estimators: (i) the unadjusted estimator $\hat{t}_{un} = N\bar{y}_r$, where $\bar{y}_r = \sum_{k\in U_r} y_k/N_r$. (ii) The one-step calibration estimator $\hat{t}_{C,1}$ given by (4.7) based on different calibration functions; see Tables 2-4.
As a measure of bias of an estimator $\hat{\theta}$ of a parameter θ, we used the Monte Carlo percent relative bias (RB)

$$RB_{MC}(\hat{\theta}) = \frac{100}{J}\sum_{j=1}^{J}\frac{(\hat{\theta}_{(j)} - \theta)}{\theta},$$

where $\hat{\theta}_{(j)}$ denotes the estimator $\hat{\theta}$ in the j-th population, $j = 1, \ldots, J$. We also computed the percent relative root mean square error (RRMSE) of $\hat{\theta}$:

$$RRMSE_{MC}(\hat{\theta}) = 100 \times \frac{\left\{J^{-1}\sum_{j=1}^{J}(\hat{\theta}_{(j)} - \theta)^2\right\}^{1/2}}{\theta}.$$

Example 1. *The y-values were generated according to the linear regression model*

$$y_k = 1\ 000 + 10x_k + \varepsilon_k,$$

where ε_k were normally distributed with mean 0 and variance 300^2. For $\hat{t}_{C,1}$, the calibration weights (4.6) were computed so that the calibration constraint, $\sum_{k\in U_r}\tilde{w}_k x_k = \sum_{k\in U} x_k$ was satisfied. Note that we have not calibrated on the population size N, which would have been a natural thing to do, if available, in this particular scenario as the relationship between y and x includes an intercept. Our goal is to show that, when (4.10) does not hold, one must rely on (4.11) to achieve unbiasedness, which implies that the calibration function is an adequate description of the inverse of the response probability.

Table (4.2) shows the RB an RRMSE (in parentheses) of several estimators. From Table (4.2), we note that the unadjusted estimator showed negligible bias under uniform nonresponse (mechanism 1), as expected. On the other hand, the calibration estimator $\hat{t}_{C,1}$, exhibited large values of RB for both linear and exponential weighting

Nonresponse mechanism	Name	ϕ_k
Example 1		
1	Uniform	0.5
2	Logistic type	$0.2 + 0.6\left\{1 + \exp(5 - x_k/8)\right\}^{-1}$
3	Inverse linear	$(1 + 0.1x_k)^{-1}$
4	Exponential	$\exp(-0.05x_k)$
Example 2		
1	Logistic type	$0.05 + 0.85\left\{1 + \exp(5 - x_k/8)\right\}^{-1}$
2	Uniform within groups	$0.6I_{(x\leq 18)} + 0.8I_{(18<x\leq 63)} + 0.3I_{(63<x)}$
Example 3		
1	Logistic type	$0.05 + 0.85\left\{1 + \exp(5 - x_k/8)\right\}^{-1}$
2	Non-monotone	$0.2 + \left\{\frac{0.6}{1 + \exp(x_k - 17)} + \frac{0.3}{1 + \exp(-x_k + 45)}\right\}$

Table 4.1: Nonresponse mechanisms used in each example

equal to -8.7% *and* -9.8%, *respectively. Assuming an uniform nonresponse mechanism,* $\phi_k = \phi_0$ *for all* k, *the approximate bias of* $\hat{t}_{C,1}$ *in (4.8) under linear weighting, reduces to*

$$\hat{t}_{C,1} \doteq (1 - \phi_0) \sum_{k \in U} (y_k - Bx_k) \neq 0,$$

in general, where $B = \sum_{k\in U} x_k y_k / \sum_{k\in U} x_k^2$. *This particular scenario clearly illustrates that performing blindly a one step calibration procedure in the presence of unit nonresponse may generate bias even when all the units have equal response probabilities. In contrast, the two-step approach would have consisted in first multiplying the design weights by the inverse of the overall response rate (after a modeling exercise would have suggested that the probability of response was unrelated to* x*) and*

performing a calibration procedure, using the adjusted weights as the starting weights.

In the case of mechanism 2, the response probabilities were generated according to a logistic type function. All the estimators exhibited some bias but, once again, the unadjusted estimator $\hat{t}_{un}$ *showed significantly less bias than* $\hat{t}_{C,1}$ *based on the linear and exponential functions (which are both inadequate descriptions of the nonresponse mechanism). As for mechanism 1, this scenario clearly illustrates that, when both (4.10) and (4.11) are not satisfied, the resulting calibration estimator is biased.*

We now turn to mechanism 3 for which the probabilities were generated according to the inverse linear function. As expected, the unadjusted estimator was biased. In contrast, the calibration estimator $\hat{t}_{C,1}$ *based on the linear function showed negligible bias, which is consistent with (4.12). The calibration estimator* $\hat{t}_{C,1}$ *based on the exponential function exhibited bias (with a RB equal to* -8.2%*), which is explained by the fact that the exponential function is not an adequate description of the nonresponse mechanism.*

Finally, under mechanism 4, units were assigned response probabilities according to the exponential function. Not surprisingly, the calibration estimator $\hat{t}_{C,1}$ *based on the exponential function showed negligible bias, whereas the linear function led to considerable bias with a RB equal to* 23.1%*.*

Nonresponse mechanism	$\hat{t}_{un}$	$\hat{t}_{C,1}$ with $F(u) = 1 + u$	$\hat{t}_{C,1}$ with $F(u) = \exp(u)$
1	-0.06	-8.7	-9.8
	(0.9)	(8.8)	(9.9)
2	6.9	-17.3	-17.5
	(7.0)	(17.3)	(17.6)
3	-10.2	0.09	-8.2
	(10.3)	(2.5)	(8.5)
4	-16.0	23.1	0.4
	(16.0)	(23.7)	(4.0)

Table 4.2: Monte Carlo percent relative and percent relative root mean square error of several estimators (in %)

Example 2. *The y-values were generated according to the quadratic linear regression model*

$$y_k = 1\ 300 + 20x_k - (x_k - 45)^2 + \varepsilon_k,$$

where ε_k *is normally distributed with mean 0 and variance* 300^2*. The calibration weights (4.6) were computed so that the calibration constraints,* $\sum_{k \in U_r} \tilde{w}_k(1, x_k) =$

$\left(N, \sum_{k\in U} x_k\right)$, *were satisfied. Unlike in Example 1, we calibrated on both the population size* N *and the population total of the* x*-values. Once again, assumption (4.10) does not hold since we omitted to include the quadratic term* x^2 *in the vector of calibration variables. Unless we are in presence of complete auxiliary information (for which the* x*-values are known for all* $k \in U$*), the population total of the* x^2*-values,* $\sum_{k\in U} x_k^2$*, is generally not available.*

In addition to the linear and exponential functions, we used the logistic function, which as been described in Kott and Liao (2012):

$$F(\boldsymbol{\lambda}^\top \boldsymbol{x}_k) = \frac{l + \exp(\boldsymbol{\gamma}^\top \boldsymbol{x}_k)}{1 + \exp(\boldsymbol{\gamma}^\top \boldsymbol{x}_k)/u}, \tag{4.14}$$

where $l = 10/9$ *and* $u = 20$ *are the lower and upper bounds of the nonresponse/calibration adjustment factors. Note that* $F(0) = 2/0.95 \neq 1$*.*

The results presented in Table 4.3 are consistent with those shown in Table 4.2. When the calibration function was chosen to be the inverse of the response probability, the resulting calibration showed negligible bias. Otherwise, it was biased. For mechanism 2, where units within a class were assigned equal probability of response, none of the calibration functions provided a correct description of the true nonresponse mechanism. As a result, the calibration estimator $\hat{t}_{C,1}$ *was biased in all the scenarios.*

Nonresponse mechanism	$\hat{t}_{un}$	$\hat{t}_{C,1}$ with $F(u) = 1 + u$	$\hat{t}_{C,1}$ with $F(u) = \exp(u)$	$\hat{t}_{C,1}$ with $F(v) = \dfrac{l + \exp(v)}{1 + \exp(v)/u}$
1	-30.2	17.8	7.5	0.2
	(30.3)	(18.0)	(7.8)	(2.1)
2	2.1	8.0	7.7	7.0
	(2.4)	(8.0)	(7.7)	(7.1)

Table 4.3: Monte Carlo percent relative and percent relative root mean square error of several estimators (in %)

Example 3. *We considered the case of a binary* y*. The* y*-values were generated according to a Bernoulli distribution with probability* ψ_k*, where*

$$\mathrm{logit}(\psi_\mathrm{k}) = 0.5(\mathrm{x_k} - 40).$$

The calibration weights (4.6) were computed so that the calibration constraints, $\sum_{k\in U_r} \tilde{w}_k(1, x_k) = \left(N, \sum_{k\in U} x_k\right)$*, were satisfied. Here, it is clear that both (4.9)*

and (4.10) do not hold. Once again, we must rely on (4.11). For mechanism 1, the logistic calibration function led to negligible bias, as expected. On the other hand, under mechanism 2 where the relationship between the response probability and x was non-monotone, all the estimators showed appreciable bias, which illustrates the difficulty of choosing an appropriate calibration function for this type of mechanism.

Nonresponse mechanism	$\hat{t}_{un}$	$\hat{t}_{C,1}$ with $F(u) = 1 + u$	$\hat{t}_{C,1}$ with $F(u) = \exp(u)$	$\hat{t}_{C,1}$ with $F(v) = \frac{l + \exp(v)}{1 + \exp(v)/u}$
1	-65.0	-11.7	-4.4	0.7
	(65.5)	(12.2)	(5.2)	(3.6)
2	1.2	7.7	7.7	7.7
	(3.2)	(7.9)	(7.8)	(7.8)

Table 4.4: Monte Carlo percent relative and percent relative root mean square error of several estimators (in %)

4.5 Discussion

Although one-step calibration is simple to implement, we are somehow reluctant to use it in multipurpose surveys. When multiple characteristics are collected, survey statisticians prefer modeling the response probability to the survey as it does not require a different model for each characteristic. In this case, complete reliance must be placed on the nonresponse model in order to achieve an efficient bias reduction for every characteristic of interest. In particular, both the choice of the explanatory variables and the form of the functional must be correctly specified, otherwise leading to potentially large biases. Our examples suggest that the induced bias may be larger than that of unadjusted estimators. In other words, the single-step approach requires a complete modeling exercise in order to ensure that both the choice of the explanatory variables and the calibration function are appropriate. For these reasons, we find the single-step approach to be a risky alternative as the choice of the calibration function is somehow "hidden". On the other hand, if one is willing to go through a model building exercise, it seems to us that the two-step approach, which separates the treatment of unit nonresponse from calibration, is attractive because it makes it possible to assess separately the impact of nonresponse and sample unbalance. Moreover, the standard practice in statistical agencies consists of adjusting the basic weights to compensate for unit nonresponse through nonparametric methods such as weighting classes based on estimated response probabilities or regression trees as both types of methods provide protection against misspecification of the

functional and account for curvature and interactions. This is especially important when the auxiliary variables are continuous and their association with the response rate is not monotonic.

Bibliography

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35(4):501–514.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133–142.

Kott, P. S. (2009). Calibration weighting: Combining probability samples and linear prediction models. *Handbook of Statistics, Sample Surveys: Inference and Analysis*, 29B:55–82.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491):1265–1275.

Little, R. J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161–168.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review / Revue Internationale de Statistique*, 54(2):pp. 139–157.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Särndal, C.-E. (2011). Three factors to signal non-response bias with applications to categorical auxiliary variables. *International Statistical Review*, 79(2):233–254.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. New York: John Wiley and Sons.

Sautory, O. (2003). Calmar 2: A new version of the calmar calibration adjustment program. In *Proceedings of the Statistics Canada Symposium*.

Chapter 5

On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse

5.1 Introduction

Weight adjustment procedures are commonly employed in surveys. The weighting approach adopted by many statistical agencies consists of two distinct steps: in the first step, the design (or basic) weights are adjusted to account for unit nonresponse. In the second step, the weights adjusted for nonresponse are further modified so that survey weighted estimates agree with known population totals. In the first step, survey statisticians aim at reducing the nonresponse bias, which may be appreciable when respondents and nonrespondents are different with respect to the survey variables. Key to achieving an efficient bias reduction is the availability of powerful auxiliary information, which is a set of variables available for both respondents and nonrespondents. At this step, the design weight of a unit is divided by its estimated response probability, which is obtained by fitting a parametric or a nonparametric nonresponse model. A frequently used method in practice consists of first dividing the respondents and nonrespondents into weighting classes and adjusting the design weights of respondents by the inverse of the response rate within each class; see, for example, Eltinge and Yansaneh (1997) and Little (1986). In the second step, some form of calibration (e.g., post-stratification) is performed in order to ensure consistency between survey weighted estimates and known population totals. Calibration procedures require auxiliary variables (called calibration variables) available for the respondents and such that their population totals are known. Moreover, if the calibration variables are related to the variables of interest, the resulting calibration estimators tend to be more efficient than the non calibrated ones.

An alternative weighting approach that has received a lot of attention recently is the so-called single step approach which uses calibration with three simultaneous goals in mind: reduce the nonresponse bias, ensure consistency between survey estimates

and known population totals and, possibly, contribute to variance reduction. Unlike the two-step approach, explicit estimation of the response probabilities is not required; see Deville (2002), Sautory (2003), Särndal and Lundström (2005), Kott (2006, 2009), Chang and Kott (2008), Kott and Chang (2010) and Kott and Liao (2012), among others. In this paper, we focus on the single step approach to weighting.

Consider a finite population U of size N. The objective is to estimate the population total $t_y = \sum_{k \in U} y_k$, of a variable of interest y. A sample, s, of size n, is selected from U according to a given sampling design $p(s)$. A complete data estimator of t_y is the expansion estimator

$$\hat{t}_\pi = \sum_{k \in s} d_k y_k,$$

where $d_k = 1/\pi_k$ denotes the design weight attached to unit k and $\pi_k = P(k \in s)$ denotes its first-order probability of inclusion in the sample. In the presence of unit nonresponse, only a subset s_r of s is observed, which makes $\hat{t}_\pi$ impossible to compute.

To define a nonresponse adjusted estimator of t_y, we assume that a vector of auxiliary variables $\mathbf{x}$ is available for $k \in s_r$ and that the vector of population totals, $\mathbf{t}_\mathbf{x} = \sum_{k \in U} \mathbf{x}_k$, is known. In practice, the $\mathbf{x}$-vector is often defined by survey managers, who wish to ensure consistency between survey weighted estimates and known population totals for some important variables (e.g., age and sex). In addition, we assume that a vector of instruments $\mathbf{z}$, of the same dimension as $\mathbf{x}$, is available for $k \in s_r$. The vector of population totals $\mathbf{t}_\mathbf{z} = \sum_{k \in U} \mathbf{z}_k$ does not need to be known. The instruments are believed to be associated with the propensity of units to respond to the survey. Let R_k be a response indicator attached to unit k such that $R_k = 1$ if unit k is a respondent and $R_k = 0$, otherwise. We consider a calibration estimator of the form

$$\hat{t}_C = \sum_{k \in s} w_k R_k y_k, \tag{5.1}$$

where

$$w_k = d_k F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{z}_k) \tag{5.2}$$

and $F(.)$ is a monotonic and twice differentiable function. The calibration weights w_k in (5.2) is the product of the design weight d_k and a weighting adjustment factor $F(\hat{\boldsymbol{\lambda}}_r^\top \mathbf{z}_k)$, which is essentially an estimate of the inverse of the response probability for unit k. Linear weighting is a special case of (5.2) for which the weights w_k are given by

$$w_k = d_k(1 + \hat{\boldsymbol{\lambda}}_r^\top \mathbf{z}_k). \tag{5.3}$$

The weights w_k in (5.2) are constructed so that the calibration constraints

$$\sum_{k \in s_r} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k \tag{5.4}$$

are satisfied. In the linear case, (5.4) implies that

$$\hat{\boldsymbol{\lambda}}_r^\top = \left(\sum_{k\in U}\mathbf{x}_k - \sum_{k\in s_r} d_k\mathbf{x}_k\right)^\top \left(\sum_{k\in s_r} d_k\mathbf{z}_k\mathbf{x}_k^\top\right)^{-1}.$$

The total error of $\hat{t}_C$ can be expressed as

$$\hat{t}_C - t_y = (\hat{t}_\pi - t_y) + (\hat{t}_C - \hat{t}_\pi). \tag{5.5}$$

The first term on the right hand side of (5.5) is the sampling error, whereas the second term is the nonresponse error. Since the sampling error does not depend on nonresponse, we focus on the nonresponse error in the sequel. Without loss of generality, we consider the case of a census $s = U$ so that the sampling error, $\hat{t}_\pi - t_y$, is equal to zero.

Regardless of the choice of the calibration function $F(.)$, the calibration estimator $\hat{t}_C$ perfectly estimates the true population total t_y if the variable of interest y is perfectly explained by the $\mathbf{x}$-vector, i.e., $y_k = \mathbf{x}_k^\top\boldsymbol{\beta}$ for some vector $\boldsymbol{\beta}$. Hence, we expect $\hat{t}_C$ to exhibit a small bias if the y-variable and the $\mathbf{x}$-vector are linearly related and the relationship is strong. However, in multipurpose surveys, the number of variables of interest is typically large (possibly few hundred) and many variables collected are categorical rather than continuous. Therefore, in most situations encountered in practice, it is unrealistic to presume that the $\mathbf{x}$-vector is linearly related to all y-variables, in which case some estimates could suffer from bias. On the other hand, if $F(\hat{\boldsymbol{\lambda}}_r^\top\mathbf{z}_k)$ is a good estimate of the inverse of the response probability of unit k, p_k^{-1}, then $\hat{t}_C$ is asymptotically unbiased for t_y regardless of the y-variable being estimated; see Section 3. For linear weighting in (5.3), Särndal and Lundström (2005, Chapter 9) showed that $\hat{t}_C$ is asymptotically unbiased for t_y for every y-variable provided that the response probability of unit k, p_k, is such that

$$p_k^{-1} = 1 + \lambda^\top\mathbf{z}_k \quad \text{for all } k \in U, \tag{5.6}$$

for a vector of unknown constants λ; see also Kott and Liao (2012) for a discussion in the case of nonlinear weighting. However, in practice, it is not clear how to select the $\mathbf{z}$-vector as these variables are available for the respondents only. Even if the $\mathbf{z}$-vector was correctly specified, it is also not clear how one would validate the form of the relationship in (5.6). The same is true for nonlinear weighting.

The purpose of this paper is to examine the so-called problem of bias amplification in the context of instrument vector calibration. In the context of epidemiological studies, it has been found that including instrumental variables in the set of conditioning variables can increase unmeasured confounding bias; see Bhattacharya and Vogt (2007), Wooldridge (2009), Pearl (2010, 2012) and Myers et al. (2011). In other words, there exists a class of auxiliary information that tends to amplify the bias if it exists. We argue that the same is true in the context of instrument vector calibration. Also, we show that, even in the absence of bias, the variance may be amplified when the calibration variables are poorly related to the instruments. Some

preliminary results in this direction can be found in Lesage (2012) and Osier (2012).

This paper is organized as follows: in Section 2, we introduce the underlying models. In Section 3, we start by examining the properties of the unadjusted estimator. Then, the properties of calibration estimators are studied and the problems of bias and variance amplification are discussed. An empirical investigation, comparing calibration estimators in terms of bias and efficiency, is conducted in Section 4. We make some final remarks in Section 5.

5.2 The underlying models

For simplicity, we consider the case of a scalar x and a scalar z. Let

$$\{(x_k, y_k, z_k, r_k)^\top, k \in \mathcal{U}\}$$

be realisations of independent and identically distributed random vectors

$$\{(X_k, Y_k, Z_k, R_k)^\top, k \in \mathcal{U}\}.$$

Without loss of generality, we assume that $\mathbb{E}(Z_k) = 0$ and $\mathbb{V}(Z_k) = 1$.

We assume that the relationship between Y and Z can be described by the model

$$\mathbb{E}(Y_k \mid Z_k) = \boldsymbol{Z}_k^\top \boldsymbol{\beta}, \tag{5.7}$$

where $\boldsymbol{Z_k} = (1, Z_k)^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. Model (5.7) is often called a prediction or outcome regression model.

We assume that

$$\mathbb{E}(R_k \mid Y_k, Z_k) = \mathbb{E}(R_k \mid Z_k) = h(Z_k; \boldsymbol{\gamma}), \tag{5.8}$$

where $\boldsymbol{\gamma}$ is a vector of unknown parameters. Model (5.8) is called the nonresponse model. We assume that (5.8) includes an intercept. The probability of response, p_k, attached to unit k is

$$p_k = \mathbb{E}\{R_k \mid (Y_k, Z_k) = (y_k, z_k)\} = h(z_k; \boldsymbol{\gamma})$$

for some known function $h(.)$. Conditionally given the realized population, the nonresponse mechanism can be described as

$$R_k \sim \mathcal{B}(1, p_k).$$

Model (5.8) implies that

$$\mathbb{C}ov(Y_k,\ R_k \mid Z_k) = 0. \tag{5.9}$$

Assumption (5.9) is essentially equivalent to MAR (Rubin, 1976). That is, we assume that there remains no residual relationship between y and the response probability after accounting for z. Figure 5.1 describes the relationship between Y, Z and R.

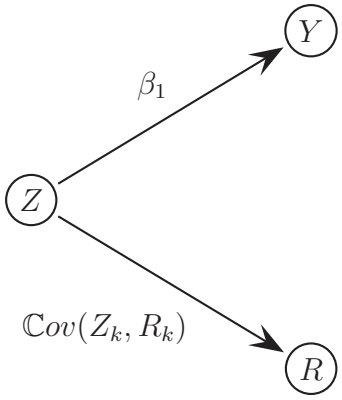


Figure 5.1: Relationship between the variables y, z and R

5.3 Properties of estimators

5.3.1 The unadjusted estimator

We start by examining the properties of the unadjusted (or naive) estimator

$$\hat{t}_{naive} = N\frac{\sum_{k\in\mathcal{U}} y_k R_k}{\sum_{k\in\mathcal{U}} R_k}. \tag{5.10}$$

A first-order Taylor expansion with respect to the nonresponse distribution leads to

$$\hat{t}_{naive} - t_y = \frac{\sum_{k\in\mathcal{U}} p_k(y_k - \bar{y}_{\mathcal{U}})}{\bar{p}_{\mathcal{U}}} + \frac{\sum_{k\in\mathcal{U}} R_k(y_k - \bar{y}_{\mathcal{U},p})}{\bar{p}_{\mathcal{U}}} + O_p(N/N_r), \tag{5.11}$$

where $\bar{y}_{\mathcal{U},p} = \sum_{k\in\mathcal{U}} p_k y_k / \sum_{k\in\mathcal{U}} p_k$, $\bar{p}_{\mathcal{U}} = \sum_{k\in\mathcal{U}} p_k/N$, $\bar{y}_{\mathcal{U}} = \sum_{k\in\mathcal{U}} y_k/N$ and N_r denotes the expected number of respondents in the population.

Ignoring the higher order terms in (5.11), the nonresponse-bias of $\hat{t}_{naive}$ can be approximated by

$$ABias_q\left(\hat{t}_{naive}\right)/N = \frac{\sum_{k\in\mathcal{U}} p_k(y_k - \bar{y}_{\mathcal{U}})/N}{\bar{p}_{\mathcal{U}}}, \tag{5.12}$$

where the subscript q denotes the assumed nonresponse model. For large N, it follows from (5.7) and (5.8) that the approximate bias in (5.12) can be approximately written as

$$\begin{aligned} ABias_q\left(\hat{t}_{naive}\right)/N &\approx \frac{\mathbb{C}ov\left(Y_k,\ R_k\right)}{\mathbb{E}(R_k)} \\ &= \beta_1 \frac{\mathbb{C}ov\left(Z_k,\ R_k\right)}{\mathbb{E}(R_k)}. \end{aligned}$$

The above expression shows that the bias of the unadjusted estimator may be large if $\beta_1\mathbb{C}ov\left(Z_k,\ R_k\right)$ is large, which in turns occurs if there is a strong association between the variable y and z and a strong association between the response status and z.

5.3.2 Properties of calibration estimators

Borrowing from the econometric literature, we start by defining the concept of proxy variable.

Definition 4. *A vector of proxy variable* $\mathbf{x}$ *for the vector of instruments* $\mathbf{z}$ *is a vector satisfying:*

1. $\mathbf{x}$ *is available for the responding units and the vector of population totals,* $t_{\mathbf{x}} = \sum_{k\in\mathcal{U}} \mathbf{x}_k$, *is known;*
2. *the system of equations*

$$\sum_{k\in\mathcal{U}} \left\{1 - R_k F\left(\boldsymbol{\lambda}^\top \boldsymbol{z}_k\right)\right\} \boldsymbol{x}_k = \mathbf{0}$$

   has a unique solution in $\boldsymbol{\lambda}$*;*

3. *the system of equations*

$$\sum_{k\in\mathcal{U}} \left\{1 - p_k F\left(\boldsymbol{\lambda}^\top \boldsymbol{z}_k\right)\right\} \boldsymbol{x}_k = \mathbf{0}$$

   has a unique solution in $\boldsymbol{\lambda}$.

Note that, in the survey sampling context, a proxy variable x corresponds to a calibration variable. Let $\boldsymbol{z}_k = (1, z_k)^\top$ be the vector of instruments and $\boldsymbol{x}_k = (1, x_k)^\top$ be the vector of calibration variables attached to unit k. We assume that the vector of population totals $\mathbf{t}_{\mathbf{x}} = (N, t_x)^\top$ is known. The estimated vector of coefficients $\hat{\boldsymbol{\lambda}}_r$ in (5.1) is defined as the solution of the sample estimating equations

$$\sum_{k\in\mathcal{U}} \left\{1 - R_k F\left(\hat{\boldsymbol{\lambda}}_r^\top \boldsymbol{z}_k\right)\right\} \boldsymbol{x}_k = \mathbf{0}.$$

Similarly, we define the vector $\boldsymbol{\lambda}_N$ as a solution of the following census estimating equations:

$$\sum_{k\in\mathcal{U}} \left\{1 - p_k F\left(\boldsymbol{\lambda}_N^\top \boldsymbol{z}_k\right)\right\} \boldsymbol{x}_k = \mathbf{0}. \tag{5.13}$$

Using mild regularity conditions (see D'Arrigo and Skinner, 2010), we have $\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_N = O_P(N_r^{-1/2})$. Using a first-order Taylor expansion, we obtain

$$\begin{aligned} F\left(\hat{\boldsymbol{\lambda}}_r^\top \boldsymbol{z}_k\right) &= F_k + f_k(\hat{\boldsymbol{\lambda}}_r - \boldsymbol{\lambda}_N)^\top \boldsymbol{z}_k + O_P(N_r^{-1}) \qquad (5.14)\\ &= F_k + \left(\boldsymbol{t}_x - \sum_{k\in\mathcal{U}} R_k F_k \boldsymbol{x}_k\right)^\top \left(\sum_{k\in\mathcal{U}} R_k f_k \boldsymbol{z}_k \boldsymbol{x}_k^\top\right)^{-1} f_k \boldsymbol{z}_k + O_P(N_r^{-1}), \end{aligned}$$

where $F_k \equiv F\left(\boldsymbol{\lambda}_N^\top \boldsymbol{z}_k\right)$ and $f_k \equiv F'\left(\boldsymbol{\lambda}_N^\top \boldsymbol{z}_k\right)$.

From (5.1) and (5.14), using Taylor expansion, it can be shown that

$$\hat{t}_C = \sum_{k\in\mathcal{U}} R_k F_k y_k + \left(\boldsymbol{t}_x - \sum_{k\in\mathcal{U}} R_k F_k \boldsymbol{x}_k\right)^\top \boldsymbol{B}_{pf} + O_P(N/N_r), \tag{5.15}$$

where

$$\boldsymbol{B}_{pf} = \left(\sum_{k\in\mathcal{U}} p_k f_k \boldsymbol{z}_k \boldsymbol{x}_k^\top\right)^{-1} \sum_{k\in\mathcal{U}} p_k f_k \boldsymbol{z}_k y_k. \tag{5.16}$$

Using (5.15), the nonresponse error of $\hat{t}_C$ can be approximated by

$$\hat{t}_C - t_y \approx -\sum_{k\in\mathcal{U}} (1 - R_k F_k)\left(y_k - \boldsymbol{x}_k^\top \boldsymbol{B}_{pf}\right). \tag{5.17}$$

The approximate bias of $\hat{t}_C$ is thus given by

$$ABias_q(\hat{t}_C) = -\sum_{k\in\mathcal{U}} (1 - p_k F_k)\left(y_k - \boldsymbol{x}_k^\top \boldsymbol{B}_{p,f}\right). \tag{5.18}$$

Now, suppose that the model linking the x-variable and the z-variable is a linear regression model given by

$$\begin{aligned} \mathbb{E}(X_k \mid Z_k) &= \boldsymbol{Z}_k^\top \boldsymbol{\alpha}, \\ \mathbb{V}(X_k \mid Z_k) &= 1 - \alpha_1^2, \end{aligned} \tag{5.19}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$. Without loss of generality, we assume that $\mathbb{V}(X_k) = 1$ and that the coefficient of correlation between x and z is equal to α_1.

Using (5.13), expression (5.18) can be rewritten as

$$\begin{aligned} ABias_q(\hat{t}_C) &= -\sum_{k\in\mathcal{U}} (1 - p_k F_k)\left\{\left(y_k - \boldsymbol{\beta}^\top \boldsymbol{z}_k\right) - \frac{\beta_1}{\alpha_1}\left(x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}\right)\right\} \\ &= -\sum_{k\in\mathcal{U}} (1 - p_k F_k)\left(y_k - \boldsymbol{z}_k^\top \boldsymbol{\beta}\right) + \frac{\beta_1}{\alpha_1} \sum_{k\in\mathcal{U}} (1 - p_k F_k)\left(x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}\right). \end{aligned} \tag{5.20}$$

From (5.20), the bias of $\hat{t}_C$ is approximately equal to zero regardless of the variable y being estimated if $p_k = F_k^{-1}$. On the other hand, the approximate bias vanishes if both model (5.7) and model (5.19) hold.

For large N, (5.20) can be approximated by

$$\begin{aligned} ABias_q(\hat{t}_C)/N &\approx \mathbb{E}\left(F_k^\infty \left[\mathbb{E}\left\{\left(Y_k - \boldsymbol{\beta}^\top \boldsymbol{Z}_k\right) R_k \mid Z_k\right\} - \frac{\beta_1}{\alpha_1}\mathbb{E}\left\{\left(X_k - \boldsymbol{\alpha}^\top \boldsymbol{Z}_k\right) R_k \mid Z_k\right\}\right]\right) \\ &\approx \mathbb{E}\left[F_k^\infty \left\{\mathbb{C}ov\left(Y_k,\ R_k \mid Z_k\right) - \frac{\beta_1}{\alpha_1}\mathbb{C}ov\left(X_k,\ R_k \mid Z_k\right)\right\}\right], \end{aligned} \tag{5.21}$$

where $F_k^{\infty} = F\left(\boldsymbol{\lambda}_{\infty}^{\top}\boldsymbol{Z}_k\right)$ and $\boldsymbol{\lambda}_{\infty}$ is the probability limit of $\boldsymbol{\lambda}_N$.
Since we have assumed (5.9), a necessary condition for the bias of $\hat{t}_C$ to vanish,

$$ABias_q(\hat{t}_C)/N \approx 0,$$

is

$$\mathbb{C}ov\left(X_k,\ R_k \mid Z_k\right) = 0. \tag{5.22}$$

That is, $\hat{t}_C$ is approximately unbiased for t_y if there remains no relationship between the proxy variable x and the response probability after accounting for the instrument z. This is consistent with the results of Kott and Chang (2010). Figure 5.2 describes the relationship between Y, Z, X and R.

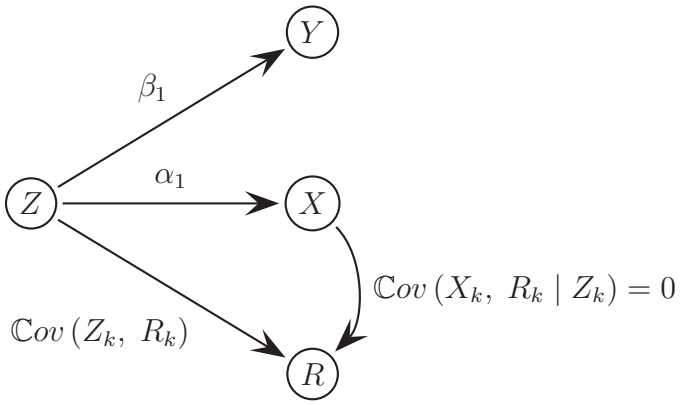


Figure 5.2: Relationship between the variables y, z, x and p

We now turn to the conditional nonresponse variance of $\hat{t}_C$. From (5.17), we obtain an approximation of the nonresponse variance:

$$\begin{aligned} A\mathbb{V}ar_q\left(\hat{t}_C\right) &= \mathbb{V}ar\left[\sum_{k\in\mathcal{U}} R_k F_k\left(y_k - \boldsymbol{x}_k^{\top}\boldsymbol{B}_{pf}\right) \mid \{(X_l, Y_l, Z_l, R_l) = (x_l, y_l, z_l, r_l)\}\right] \\ &= \sum_{k\in\mathcal{U}} (1-p_k) p_k F_k^2\left(y_k - \boldsymbol{x}_k^{\top}\boldsymbol{B}_{pf}\right)^2. \end{aligned} \tag{5.23}$$

It is shown in the Appendix that (5.23) can alternatively be written as

$$A\mathbb{V}ar_q\left(\hat{t}_C\right) = \sum_{k\in\mathcal{U}} (1-p_k) p_k F_k^2 \left\{\left(y_k - \boldsymbol{z}_k^{\top}\hat{\boldsymbol{\beta}}_{pf}\right) - \frac{\hat{\beta}_{pf,1}}{\hat{\alpha}_{pf,1}}\left(x_k - \boldsymbol{z}_k^{\top}\hat{\boldsymbol{\alpha}}_{pf}\right)\right\}^2 \tag{5.24}$$

where

$$\hat{\boldsymbol{\beta}}_{pf} = (\hat{\beta}_{pf,0}, \hat{\beta}_{pf,1})^{\top} = \left(\sum_{k\in\mathcal{U}} p_k f_k \boldsymbol{z}_k \boldsymbol{z}_k^{\top}\right)^{-1} \sum_{k\in\mathcal{U}} p_k f_k \boldsymbol{z}_k y_k$$

and

$$\hat{\boldsymbol{\alpha}}_{pf} = (\hat{\alpha}_{pf,0}, \hat{\alpha}_{pf,1})^{\top} = \left(\sum_{k\in\mathcal{U}} p_k f_k \boldsymbol{z}_k \boldsymbol{z}_k^{\top}\right)^{-1} \sum_{k\in\mathcal{U}} p_k f_k \boldsymbol{z}_k x_k. \tag{5.25}$$

From (5.24), the variance of $\hat{t}_C$ is small if the residuals $(y_k - \boldsymbol{z}_k^\top \hat{\boldsymbol{\beta}}_{pf})$ and $(x_k - \boldsymbol{z}_k^\top \hat{\boldsymbol{\beta}}_{pf})$ are small and $\hat{\alpha}_{pf,1}$ is large (which corresponds to a strong association between x and z). Therefore, a small value of $\hat{\alpha}_{pf,1}$ (i.e., a low correlation between x and z), which corresponds to a weak proxy x, may translate into a highly unstable calibration estimator. In this case, we are in the presence of variance amplification. This is illustrated in the empirical study presented in Section 4; see also Osier (2012).

5.3.3 The problem of bias amplification

In Section 3.2, we argued that the calibration estimator $\hat{t}_C$ is approximately unbiased provided that $\mathbb{C}ov(R_k, X_k \mid Z_k) = 0$. Here, we examine the situation, where $\mathbb{C}ov(R_k, X_k \mid Z_k) \neq 0$. To that end, assume that there exists an unobserved variable u, independent of z and y, which is related to both the response indicator variable r and the proxy variable x; see Figure (5.3). The random variables U_k are part of the superpopulation model and we now consider $\{(x_k, y_k, z_k, u_k, r_k)^\top, k \in \mathcal{U}\}$ as realisations of independent and identically distributed random vectors $\{(X_k, Y_k, Z_k, U_k, R_k)^\top, k \in \mathcal{U}\}$. Without loss of generality, we assume that $\mathbb{E}(U_k \mid Z_k) = 0$ and $\mathbb{V}(U_k \mid Z_k) = 1$.

The relation between r and u implies that

$$\mathbb{C}ov\left(U_k,\ R_k \mid Z_k\right) \neq 0. \tag{5.26}$$

Now, suppose that the variables x and u are related through

$$\begin{aligned} \mathbb{E}\left(X_k \mid Z_k, U_k\right) &= \alpha_0 + \alpha_1 Z_k + \alpha_2 U_k, \\ \mathbb{V}\left(X_k \mid Z_k, U_k\right) &= 1 - \alpha_1^2 - \alpha_2^2. \end{aligned} \tag{5.27}$$

Since $\alpha_2 \neq 0$, there is a linear relationship between the proxy variable x and the unobserved variable u. From (5.26) and (5.27), it follows that

$$\mathbb{C}ov(X_k,\ R_k \mid Z_k) = \alpha_2\, \mathbb{C}ov\left(U_k,\ R_k \mid Z_k\right) \neq 0.$$

The relationship between y and z is still described by

$$\mathbb{E}(Y_k \mid U_k, Z_k) = \beta_0 + \beta_1 Z_k. \tag{5.28}$$

it follows from (5.28) that $\mathbb{C}ov(Y_k,\ U_k \mid Z_k) = 0$ and it follows that we still have

$$\mathbb{C}ov(Y_k,\ R_k \mid Z_k) = 0.$$

Théoreme 1. *If $\mathbb{C}ov(X_k, R_k \mid Z_k) \neq 0$, then expression of the bias (5.21) becomes:*

$$ABias_q(\hat{t}_c) \ \approx\ -\beta_1 \frac{\alpha_2}{\alpha_1} \mathbb{E}\left\{F_k^\infty\, \mathbb{C}ov\left(U_k,\ R_k \mid Z_k\right)\right\}. \tag{5.29}$$

Expression (5.29) suggests $\hat{t}_C$ is biased if $\mathbb{C}ov\,(X_k,\ R_k \mid Z_k) = \alpha_2\ \mathbb{C}ov\,(U_k,\ R_k \mid Z_k) \neq 0$. The bias is large if α_2 is large (i.e., there is a strong association between the u-variable and the proxy variable x) and/or if $\mathbb{C}ov\,(U_k,\ R_k \mid Z_k)$ is large (i.e., there is a strong association between the u-variable and the probability of response). For a given value of $\alpha_2\ \mathbb{C}ov\,(U_k,\ R_k \mid Z_k)$, the bias is amplified if α_1 is small; i.e., if the relationship between x and z is weak. Therefore, it may be wise to select x-variables that are strongly correlated with the instrument z as it would help preventing from bias amplification. However, as mentioned in Section 1, the proxy variables corresponds to the calibration variables, which are often defined by survey managers. For example, the variables age and sex are often part of the vector of calibration variables in household and social surveys in order to ensure consistency between survey weighted estimates and their known population counts. If these variables are poorly related to the instrument z, they may contribute in significantly increasing the bias of the calibration estimator $\hat{t}_C$. This is illustrated in the empirical study presented in the next section.

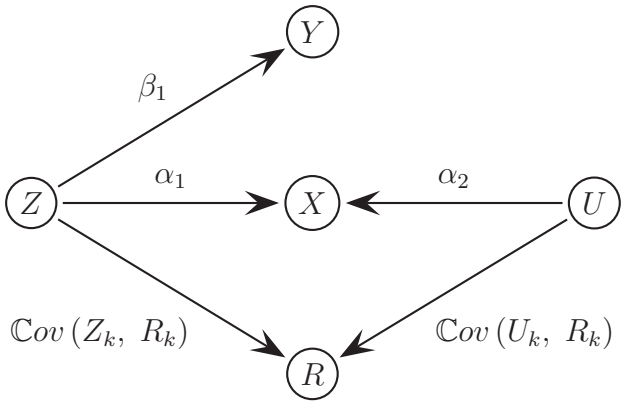


Figure 5.3: Relationship between the variables y, z, x, u and r

5.4 Simulation study

We conducted two simulation studies in order to illustrate the problem of bias and variance amplification.

5.4.1 Simulation study 1

We generated finite populations of size $N = 1\ 000$, each consisting of a variable of interest y, a proxy variable $x^{(\alpha_1,\alpha_2)}$ with $\alpha_1 \in \{0.2, 0.3, 0.5, 0.7\}$ and $\alpha_2 \in \{0, 0.1, 0.3, 0.5\}$, an instrumental variable z and an unobserved variable u.

First, the variables z and u were generated from a uniform distribution $\left(-\sqrt{3}, \sqrt{3}\right)$ so that $\mathbb{E}(Z) = \mathbb{E}(U) = 0$ and $\mathbb{V}(Z) = \mathbb{V}(U) = 1$. Then, given the z-values, the y-values were generated according to the linear regression model

$$Y_k = 10 + 2z_k + \varepsilon_k^y,$$

where ε_k^y is normally distributed with mean 0 and variance 1. The resulting coefficient of determination was equal to 79.2%.

Finally, the $x^{(\alpha_1,\alpha_2)}$-values were generated according to the linear regression model

$$X_k^{(\alpha_1,\alpha_2)} = \alpha_1 z_k + \alpha_2 u_k + \sigma_{(\alpha_1,\alpha_2)} \varepsilon_k^{(\alpha_1,\alpha_2)},$$

where $\sigma^2_{(\alpha_1,\alpha_2)} = 1 - \alpha_1^2 - \alpha_2^2$ and the errors $\varepsilon^{(\alpha_1,\alpha_2)}$ were normally distributed with mean equal to 0 and variance equal to 1.

In order to focus on the nonresponse error, we considered the census case; i.e., $n = N = 1\ 000$. In each population, units were assigned a response probability p_k according to

$$\text{logit}(p_k) = 1.5 z_k + u_k.$$

The overall response rate was set to 50% approximately. Finally, the response indicators R_k were generated independently from a Bernoulli distribution with parameter p_k, $k \in U$.

The whole process (i.e., generating the finite population and generating nonresponse), was repeated $K = 10,000$ times, leading to $K = 10,000$ sets of respondents.

In each sample containing respondents and nonrespondents, we computed the following estimators: (i) the naive estimator, $\hat{t}_{naive}$, given by (5.10); (ii) the instrumental calibration estimator, $\hat{t}_C$ given by (5.1) with linear weighting defined in (5.3) and two calibration totals N and $t_{x^{(\alpha_1,\alpha_2)}}$. We note by $\hat{t}_C(\alpha_1,\alpha_2)$ the estimator $\hat{t}_C$ obtained for given values of α_1 and α_2.

For an estimator $\hat{t}$, we computed the Monte Carlo percent relative bias given by

$$RB_{MC}\left(\hat{t}\right) = \frac{1}{K}\sum_{j=1}^{K} \frac{(\hat{t}_{(j)} - t_{y(j)})}{t_{y(j)}} \times 100,$$

where $\hat{t}_{(j)}$ denotes the estimator $\hat{t}$ and $t_{y(j)}$ denotes the true population total t_y at the j-th iteration.

As a measure of variability of $\hat{t}$, we computed the Monte Carlo coefficient of variation (in %) given by

$$CV_{MC}(\hat{t}) = 100 \times \frac{\left[\frac{1}{K}\sum_{j=1}^{K}\left(\hat{t}_{(j)} - t_{y(j)}\right)^2 - \left\{E_{MC}(\hat{t} - t_y)\right\}^2\right]^{0.5}}{E_{MC}(t_y)},$$

where $E_{MC}(\hat{t}) = \sum_{j=1}^{K} \hat{t}_{(j)}/K$ and $E_{MC}(t_y) = \sum_{j=1}^{K} t_{y(j)}/K$. Table 5.1 shows the relative bias and the coefficient of variation (in parentheses) of $\hat{t}_C(\alpha_1,\alpha_2)$.

Before discussing the results shown in Table 1, it is worth noting that the naive estimator showed values of relative bias and coefficient of variation equal to 8.3% and 0.7%, respectively.

From Table 5.1, we note that the calibration estimator $\hat{t}_C$ showed negligible bias when $\alpha_2 = 0$. These results are not surprising since $\alpha_2 \mathbb{C}ov\,(X_k,\ p_k \mid Z_k) = 0$ when $\alpha_2 = 0$; see Section 3. In this case, the variance of $\hat{t}_C$ increased rapidly as α_1 decreased. For example, for $\alpha_1 = 0.7$, the coefficient of variation of $\hat{t}_C$ was equal to 0.9%, whereas it was equal to 4.6% for $\alpha_1 = 0.2$. Similar results were reported in Osier (2012). We now turn to the case of $\alpha_2 > 0$. In this case, we have $\alpha_2\ \mathbb{C}ov\,(X_k,\ R_k \mid Z_k) \neq 0$ so the calibration estimator $\hat{t}_C$ exhibited some bias. The bias increased as α_2 increased. For a given value of $\alpha_2 > 0$, both the bias and the variance of $\hat{t}_C$ increased rapidly, which clearly illustrates that it suffers simultaneously from bias and variance amplification. For example, for $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$ the relative bias and coefficient of variation were equal to 2.8% and 1%, respectively, whereas they were equal to -13.8% and 61.9% for $\alpha_1 = 0.2$ and $\alpha_2 = 0.3$. Finally, the naive estimator performed better than $\hat{t}_C$ for large values of α_1 and α_2. For example, for $\alpha_1 = 0.3$ and $\alpha_2 = 0.5$, the calibration estimator $\hat{t}_C$ showed values of relative bias and coefficient of variation equal to -14% and 5.9%, respectively, which were significantly larger than that of the naive estimator. Figure 5.4 shows side by side boxplots of the relative error of $\hat{t}_C$, $100 \times (\hat{t}_C - t_y)/t_y$, each boxplot corresponding to the calibration estimator $\hat{t}_C(\alpha_1, \alpha_2)$ for a given pair (α_1, α_2).

α_1 \ α_2	0	0.1	0.3	0.5
0.7	0.02	-0.9	-2.8	-4.9
	(0.9)	(0.9)	(1.0)	(1.1)
0.5	-0.1	-1.3	-4.1	-7.2
	(1.4)	(1.5)	(1.7)	(2.1)
0.3	-0.2	-2.4	-7.5	-14.0
	(2.6)	(3.0)	(4.1)	(5.9)
0.2	-0.6	-4.5	-13.8	-27.4
	(4.6)	(15.6)	(61.9)	(65.6)

Table 5.1: Monte Carlo percent relative bias and percent coefficient of variation (in parentheses) of $\hat{t}_C$ for different pairs (α_1, α_2).

5.4.2 Simulation study 2

We used the same models as in Section 4.1, except that the instrument z now coincides with the variable of interest y; see Figure 5.5. The proxy variable was

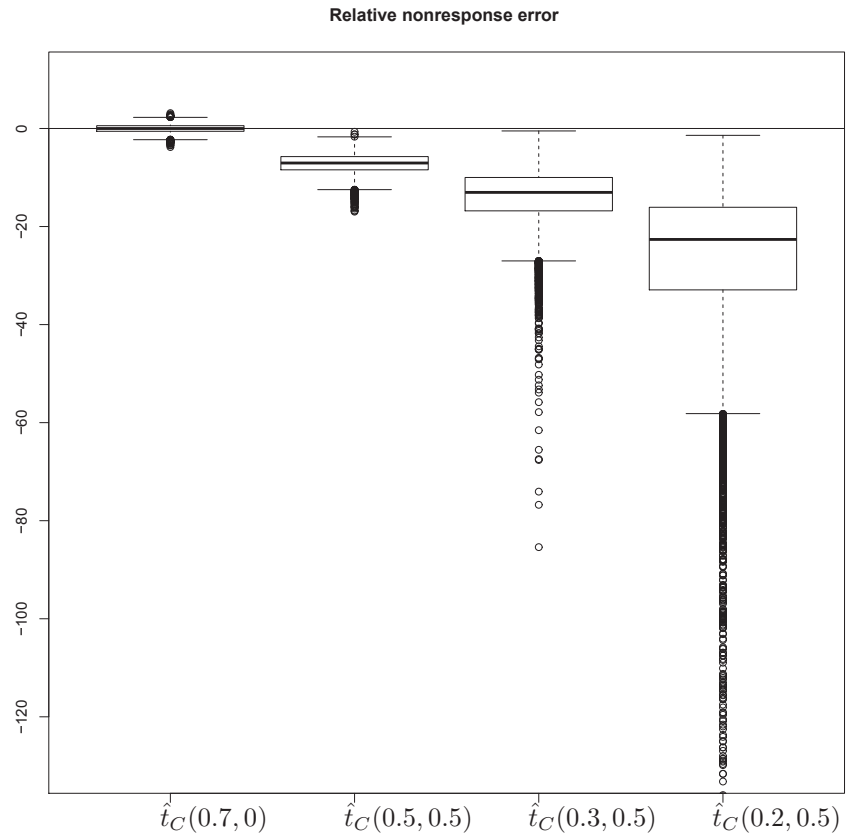


Figure 5.4: Boxplot of relative errors (in %) for different pairs (α_1, α_2)

generated according to

$$X_k^{(\alpha_1,\alpha_2)} = \alpha_1 \mathbb{V}ar(y)^{-1}(y_k - \beta_0) + \alpha_2 u_k + \sigma_{(\alpha_1,\alpha_2)}\varepsilon_k^{(\alpha_1,\alpha_2)},$$

where $\sigma^2_{(\alpha_1,\alpha_2)} = 1 - \alpha_1^2 - \alpha_2^2$ and the errors $\varepsilon^{(\alpha_1,\alpha_2)}$ were normally distributed with mean equal to 0 and variance equal to 1. Units were assigned response probabilities, p_k, such that

$$\text{logit}(p_k) = -5 + 0.5y_k + u_k.$$

In this study, there were a causal relationship between the response probability and the variable of interest y. Once again, the overall response rate was set to 50% approximately.

In each sample, we computed the calibration estimator in (5.1) with $z = y$. Table 5.2 shows the Monte Carlo percent relative bias and the Monte Carlo coefficient of variation (in parentheses) of $\hat{t}_C(\alpha_1, \alpha_2)$. From Table 5.2, we note that the results were very similar to those obtained in Section 4.1. Once again, $\hat{t}_C$ showed negligible bias for $\alpha_2 = 0$. In this particular scenario, it is worth noting that instrument vector calibration was successful in eliminating the nonresponse bias, despite the causal relationship between the response probability and the variable of interest. Nevertheless, it suffered from instability for small values of α_1 as the variance increased

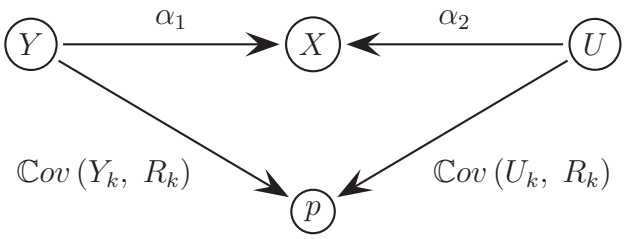


Figure 5.5: Relationship between the variables y, z, x, u and r

rapidly as α_1 decreased. For a given value of $\alpha_2 > 0$, both the bias and the variance of $\hat{t}_C$ increased as α_1 decreased.
The naive estimator showed values of relative bias and coefficient of variation equal to 7.2% and 0.7%, respectively. Once again, the latter performed better than $\hat{t}_C$ for large values of α_1 and α_2.

α_1 \ α_2	0	0.1	0.3	0.5
0.7	-0.01	-1.0	-3.3	-5.7
	(0.8)	(0.8)	(0.9)	(1.0)
0.5	-0.1	-1.5	-4.8	-8.3
	(1.4)	(1.5)	(1.7)	(2.0)
0.3	-0.2	-2.7	-8.7	-15.5
	(2.6)	(2.9)	(4.0)	(5.5)
0.2	-0.5	-4.7	-15.0	-30.2
	(5.0)	(7.1)	(25.2)	(39.1)

Table 5.2: Monte Carlo percent relative bias and percent coefficient of variation (in parentheses) of $\hat{t}_C$ for different pairs (α_1, α_2).

5.5 Discussion

In this paper, we showed that instrument vector calibration may be successful in reducing the nonresponse bias. However, in some situations, this type of calibration procedure may result in highly biased and/or unstable estimators. We first showed that instrument vector calibration leads to negligible bias provided that $\mathbb{C}ov\,(X_k,\ R_k \mid Z_k) = 0$ but that the resulting estimator may be very unstable if the calibration (or proxy) variables are weakly related to the instruments. Ideally, the calibration variables should be those exhibiting a strong relationship

with the instruments, which may require some kind of statistical modeling. When $\mathbb{C}ov\left(X_k,\ R_k \mid Z_k\right) \neq 0$, we showed that the calibration estimator is biased. Both the bias and the variance are amplified as the relationship between the calibration variables and the instruments gets weaker.

Alternatively, one may use the usual calibration based solely on calibration variables for which the population total is known. Although one may not be successful in reducing the bias to the same extent as with instrument vector calibration, there is no risk of bias and variance amplification as the calibration variables coincide with the instruments. As a result, the relationship between the calibration variables and the instruments is perfect, which in turns prevents from obtaining a point estimator with an unduly large bias and/or variance.

Bibliography

Bhattacharya, J. and Vogt, W. B. (2012). Do instrumental variables belong in propensity scores? *International Journal of Statistics & Economics*, 9(A12):107–127.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95(3):555–571.

D'Arrigo, J. and Skinner, C. J. (2010). Linearization variance estimation for generalized raking estimators in the presence of nonresponse. *Survey methodology*, 36:181–192.

Deville, J. (2002). La correction de la non-réponse par calage généralisé. *Actes des journées de méthodologie statistique, INSEE*, pages 4–20.

El Tinge, J. and Yansaneh, I. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the us consumer expenditure survey. *Survey methodology*, 23:33–40.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133–142.

Kott, P. S. (2009). Calibration weighting: Combining probability samples and linear prediction models. *Handbook of Statistics, Sample Surveys: Inference and Analysis*, 29B:55–82.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491):1265–1275.

Kott, P. S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. In *Survey Research Methods*, volume 6, pages 105–111.

Lesage, E. Correction de la non-réponse non ignorable par une approche modèle. In *Actes des Journées de Méthodologie Statistique de l'INSEE*.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review / Revue Internationale de Statistique*, 54(2):pp. 139–157.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–1222.

Osier, G. Traitement de la non-réponse non-ignorable par calage généralisé : une simulation à partir de l'enquête budget des ménages au luxembourg. In *Actes des Journées de Méthodologie Statistique de l'INSEE*.

Pearl, J. (2012). On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. New York: John Wiley and Sons.

Sautory, O. (2003). Calmar 2: A new version of the calmar calibration adjustment program. In *Proceedings of the Statistics Canada Symposium*.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables. *Unpublished Manuscript]. East Lansing, MI: Michigan State University*.

Appendix

We derive expression (5.24). We start by writing model (5.19) as

$$\boldsymbol{X}_k = A^\top \boldsymbol{Z}_k + \begin{pmatrix} 0 \\ \varepsilon_k^x \end{pmatrix}, \tag{5.30}$$

where

$$\boldsymbol{X}_k = (1, X_k)^\top,$$

$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha_0 \\ 0 & \alpha_1 \end{pmatrix}$$

and

$$\mathbb{E}(\varepsilon_k^x \mid \boldsymbol{Z}_k) = 0.$$

It follows that

$$\hat{\boldsymbol{A}}_{pf} = \begin{pmatrix} 1 & \hat{\alpha}_{pf,0} \\ 0 & \hat{\alpha}_{pf,1} \end{pmatrix} = \left(\sum_{k \in \mathcal{U}} p_k f_k \boldsymbol{z}_k \boldsymbol{z}_k^\top \right)^{-1} \sum_{k \in \mathcal{U}} p_k f_k \boldsymbol{z}_k \boldsymbol{x}_k^\top. \tag{5.31}$$

From (5.25), (5.31) and (5.16), we obtain

$$\hat{\boldsymbol{\beta}}_{pf} = \hat{\boldsymbol{A}}_{pf} \boldsymbol{B}_{pf}, \tag{5.32}$$

which leads to

$$\boldsymbol{B}_{pf} = \begin{pmatrix} \left(\hat{\alpha}_{pf,1} \hat{\beta}_{pf,0} - \hat{\alpha}_{pf,0} \hat{\beta}_{pf,1} \right) / \hat{\alpha}_{pf,1} \\ \hat{\beta}_{pf,1} / \hat{\alpha}_{pf,1} \end{pmatrix}. \tag{5.33}$$

It follows from (5.32) that

$$y_k - \boldsymbol{B}_{pf}^\top \boldsymbol{x}_k = (y_k - \hat{\boldsymbol{\beta}}_{pf}^\top \boldsymbol{z}_k) - \boldsymbol{B}_{pf}^\top (\boldsymbol{x}_k - \hat{\boldsymbol{A}}_{pf}^\top \boldsymbol{z}_k).$$

Since $\boldsymbol{x}_k - \hat{\boldsymbol{A}}_{pf}^\top \boldsymbol{z}_k = \begin{pmatrix} 0 \\ x_k - \hat{\boldsymbol{\alpha}}_{pf}^\top \boldsymbol{z}_k \end{pmatrix}$, we obtain

$$\boldsymbol{B}_{pf}^\top (\boldsymbol{x}_k - \boldsymbol{A}_{pf}^\top \boldsymbol{z}_k) = \frac{\hat{\beta}_{pf,1}}{\hat{\alpha}_{pf,1}} (x_k - \boldsymbol{\alpha}_{pf}^\top \boldsymbol{z}_k),$$

which proves expression (5.24).

Chapter 6

Conditional inference with a complex sampling: exact computations and Monte Carlo estimations

6.1 Introduction

The purpose of this article is to give a systematic use of the auxiliary information at the estimation phase by the means of Monte Carlo methods, in a design based approach.

In survey sampling, we often face a situation where we use information about the population (auxiliary information) available only at the estimation phase. For example, this information can be provided by an administration file available only posterior to the collection stage. Another example would be the number of respondents to a survey. It is classical to deal with the non-response mechanism by a second sampling phase (often Poisson sampling conditional to the size of the sample). The size of the respondents sample is known only after the collection.

This information can be compared to its counterpart estimated by the means of the sample. A significant difference typically reveals an unbalanced sample. In order to take this discrepancy into account, it is necessary to re-evaluate our estimations. In practice, two main technics exist: the model-assisted approach (ratio estimator, post-stratification estimator, regression estimator) and the calibration approach. The conditional approach we will develop in this chapter has been so far mainly a theoretical concept because it involves rather complex computations of the inclusion probabilities. The use of Monte-Carlo methods could be a novelty that would enable the use of conditional approach in practice. In particular, it seems to be very helpful for the treatment of outliers and strata jumpers.

Conditional inference in survey sampling means that, at the estimation phase, the sample selection is modelized by means of a conditional probability. Hence, expectation and variance of the estimators are computed according to this conditional sampling probability. Moreover, we are thus provided with conditional sampling

weights with better properties than the original sampling weights, in the sense that they lead to a better balanced sample (or calibrated sample).

Conditional inference is not a new topic and several authors have studied the conditional expectation and variance of estimators, among them: Rao (1985), Robinson(1987), Tillé (1998 et 1999) and Andersson (2004). Moreover, one can see that the problematic of conditional inference is close to inference in the context of rejective sampling design. The difference is that in rejective sampling, the conditioning event is controlled by the design, whereas, in conditional inference, the realization of the event is observed.

In section 2, the classical framework of finite population sampling and some notations are presented.

In section 3, we discuss the well-known setting of simple random sampling where we condition on the sizes of the sub-samples on strata (a posteriori stratification). This leads to an alternative estimator to the classical HT estimator. While a large part of the literature deals with the notion of correction of conditional bias, we will directly use the concept of conditional HT estimator (Tillé, 1998), which seems more natural under conditional inference. A simulation study will be performed in order to compare the accuracy of the conditional strategy to the traditional one.

In section 4, the sampling design is a Poisson sampling conditional to sample size n (also called conditional Poisson sampling of size n). We use again the information about the sub-samples sizes to condition on. We show that the conditional probability corresponds exactly to a stratified conditional Poisson sampling and we give recursive formula that enables the calculation of the conditional inclusion probabilities.

In section 5, we use a new conditioning statistic. Following Tillé (1998 et 1999), we use the non-conditional HT estimation of the mean of the auxiliary variable to condition on. Whereas Tillé uses asymptotical arguments in order to approximate the conditional inclusion probabilities, we prefer to perform Monte Carlo simulations to address a non-asymptotic setting. Note that this idea of using independent replications of the sampling scheme in order to estimate inclusion probabilities when the sampling design is complex has been already proposed by Fattorini (2006) and Thompson and Wu (2008).

In section 6, we apply this method to practical examples: outlier and strata jumper in business survey. This new method to deal with outliers gives good results.

6.2 The context

Let U be a finite population of size N. The statistical units of the population are indexed by a label $k \in \{1, ..., N\}$. A random sample without replacement s

is selected using a probability (sampling design) $p(.)$. $\mathcal{S}$ is the set of the possible samples s. $I_{[k\in s]}$ is the indicator variable which is equal to one when the unit k is in the sample and 0 otherwise. The size of the sample is $n(s) = |s|$. Let $B_k = \{s \in \mathcal{S}, k \in s\} = \{s \in \mathcal{S}, I_{[k\in s]} = 1\}$ be the set of samples that contain k. For a fixed individual k, let $\pi_k = p(B_k)$ be the inclusion probability and let $d_k = \frac{1}{\pi_k}$ be its sampling weight. For any variable z that takes the value z_k on the U-unit k, the sum $t_z = \sum_{k\in U} z_k$ is referred to as the total of z over U. $\widehat{t}_{z,\pi} = \sum_{k\in s} \frac{1}{\pi_k} z_k$ is the Horvitz-Thompson estimator of the total t_z.

Let x be an auxiliary variable that takes the value x_k for the individual k. The x_k are assumed to be known for all the units of U. Such auxiliary information is often used at the sampling stage in order to improve the sampling design. For example, if the auxiliary variable is a categorical variable then the sampling can be stratified. If the auxiliary variable is quantitative, looking for a balanced sampling on the total of x is a natural idea. These methods reduce the size of the initial set of admissible samples. In the second example, $\mathcal{S}_{balanced} = \{s \in \mathcal{S}, \widehat{t}_{x,\pi} = t_x\}$.
We wish to use auxiliary information *after* the sample selection, that is to take advantage of information such as the number of units sampled in each stratum or the estimation of the total t_x given by the Horvitz-Thompson estimator. Let us take an example where the sample consists in 20 men and 80 women, drawn by a simple random sampling of size $n = 100$ among a total population of $N = 200$ with equal inclusion probabilities $\pi_k = 0.5$. And let us assume that we are given *a posteriori* the additional information that the population has 100 men and 100 women. Then it is hard to maintain anymore that the inclusion probability for both men and women was actually 0.5. It seems more sensible to consider that the men sampled had indeed a inclusion probability of 0.2 and a weight of 5. Conditional inference aims at giving some theoretical support to such feelings.

We use the notation $\mathbf{\Phi}(s)$ for the statistic that will be used in the conditioning. $\mathbf{\Phi}(s)$ is a random vector that takes values in $\mathbb{R}^q$. In fact, $\mathbf{\Phi}(s)$ will often be a discrete random vector which takes values in $\{1, ..., n\}^q$. At each possible subset $\varphi \subset \mathbf{\Phi}(\mathcal{S})$ corresponds an event $A_\varphi = \mathbf{\Phi}^{-1}(\varphi) = \{s \in \mathcal{S}, \mathbf{\Phi}(s) \in \varphi\}$.

For example, if the auxiliary variable x_k is the indicator function of a domain, say $x_k = 1$ if the unit k is a man, then we can choose $\mathbf{\Phi}(s) = \sum_{k\in s} I_{[k\in domain]} = n_{domain}$ the sample size in the domain (number of men in the sample). If the auxiliary variable x_k is a quantitative variable, then we can choose $\mathbf{\Phi}(s) = \sum_{k\in s} \frac{x_k}{\pi_k} = \widehat{t}_{x,\pi}$ the Horvitz-Thompson estimator of the total t_x.

6.3 A Posteriori Simple Random Sampling Stratification

6.3.1 Classical Inference

In this section, the sampling design is a simple random sampling without replacement(SRS) of fixed size n; $\mathcal{S}_{SRS} = \{s \in \mathcal{S}, n(s) = n\}$; $p(s) = 1/\binom{N}{n}$ and the inclusion probability of each individual k is $\pi_k = n/N$. Let y be the variable of study. y takes the value y_k for the individual k. The y_k are observed for all the units of the sample. The Horvitz-Thompson (HT) estimator of the total $t_y = \sum_{k\in U} y_k$ is $\hat{t}_{y,HT} = \sum_{k\in U} \frac{y_k}{\pi_k} I_{[k\in s]}$.

Assume now that the population U is split into H sub-populations U_h called strata. Let $N_h = |U_h|$, $h \in \{1, ..., H\}$ be the auxiliary information to be taken into account. We split the sample s into H sub-samples s_h defined by $s_h = s \cap U_h$. Let $n_h(s) = |s_h|$ be the size of the sub-sample s_h.

Ideally, to use the auxiliary information at the sampling stage would be best. Here, a simple random stratified sampling (SRS stratified) with a proportional allocation $N_h n/N$ would be more efficient than a SRS. For such a SRS stratified, the set of admissible samples is $\mathcal{S}_{SRSstratified} = \{s \in \mathcal{S}, \forall h \in [1, H], n_h(s) = N_h n/N\}$, and the sampling design is $p(s) = \prod_{h\in[1,H]} \frac{1}{\binom{N_h}{n_h}}, \quad s \in \mathcal{S}_{SRSstratified}$. Once again, our point is precisely to consider setting where the auxiliary information becomes available *posterior* to this sampling stage

6.3.2 Conditional Inference

The *a posteriori* stratification with an initial SRS was described by Rao(1985) and Tillé((Tillé, 1998)). A sample s_0 of size $n(s_0) = n$ is selected. We observe the sizes of the strata sub-samples: $n_h(s_0) = \sum_{k\in U_h} I_{[k\in s]}$, $h \in [1, H]$. We assume that $\forall h, n_h(s_0) > 0$. We then consider the event:

$$A_0 = \{s \in \mathcal{S}, \forall h \in [1, H], n_h(s) = n_h(s_0)\}.$$

It is clear that $s_0 \in A_0$, so A_0 is not empty.

We consider now the conditional probability: $p^{A_0}(.) = p(./A_0)$ which will be used as far inference is concerned. The conditional inclusion probabilities are denoted

$$\pi_k^{A_0} = p^{A_0}\left([I_{[k\in s]} = 1]\right) = \mathbb{E}^{A_0}\left(I_{[k\in s]}\right) = p\left([I_{[k\in s]} = 1] \cap A_0\right)/p(A_0).$$

Accordingly, we define the conditional sampling weights: $d_k^{A_0} = \frac{1}{\pi_k^{A_0}}$.

Proposition 7. *1. The conditional probability p^{A_0} is the law of a stratified simple random sampling with allocation $(n_1(s_0), ..., n_H(s_0))$,*

2. For a unit k of the strata h: $\pi_k^{A_0} = \dfrac{n_h(s_0)}{N_h}$ *and* $d_k^{A_0} = \dfrac{N_h}{n_h(s_0)}$.

Proof. $|A_0| = \binom{N_1}{n_1(s_0)} \times ... \times \binom{N_H}{n_H(s_0)}$.
$\forall s \in A_0$, $p^{A_0}(s) = 1/|A_0|$. So we have:

$$\begin{aligned} p^{A_0}(s) &= I_{[s\in A_0]} \frac{1}{\prod_{h\in[1,H]} \binom{N_h}{n_h(s_0)}} \\ &= I_{[s\in A_0]} * \prod_{h\in[1,H]} \frac{1}{\binom{N_h}{n_h(s_0)}} \\ &= \prod_{h\in[1,H]} I_{[n_h(s)=n_h(s_0)]} * \frac{1}{\binom{N_h}{n_h(s_0)}} \end{aligned}$$

and we recognize the probability law of a stratified simple random sampling with allocation $(n_1(s_0), ..., n_H(s_0))$.
2. follows immediately. □

Note that

$$\mathbb{E}^{A_0}\left(\sum_{k\in U} \frac{y_k}{\pi_k} I_{[k\in s]}\right) = \sum_{k\in U} \frac{y_k}{\pi_k} \pi_k^{A_0} = \sum_h \sum_{k\in U_h} y_k \frac{N n_h(s_0)}{n N_h},$$

so that the genuine HT estimator is conditionally biased in this framework.
Even if, as Tillé((Tillé, 1998)) mentioned, it is possible to correct this bias simply by retrieving it from the HT estimator, it seems more coherent to use another linear estimator constructed like the HT estimator but, this time, using the conditional inclusion probabilities.

Remark that in practice A_0 should not be too small. The idea is that for any unit k, we should be able to find a sample s such that $s \in A_0$ and $k \in s$. Thus, all the units of U have a positive conditional inclusion probability.

Definition 5. *The* ***conditional HT estimator*** *is defined as:*

$$\hat{t}_{y,CHT} = \sum_{k\in U} \frac{y_k}{\pi_k^{A_0}} I_{[k\in s]}$$

The conditional Horvitz-Thompson (CHT) etimator is obviously conditionally unbiased and, therefore, unconditionally unbiased.

This estimator is in fact the classical post-stratification estimator obtained from a model-assisted approach (see (Särndal et al., 1992) pages 264-269 for example). However, conditional inference leads to a different derivation of the variance, which appears to be more reliable as we will see in next subsection.

6.3.3 Simulations

In this part, we will compare the punctual estimations of a total according to two strategies: (SRS design + conditional (post-stratification) estimator) and (SRS design + HT estimator).

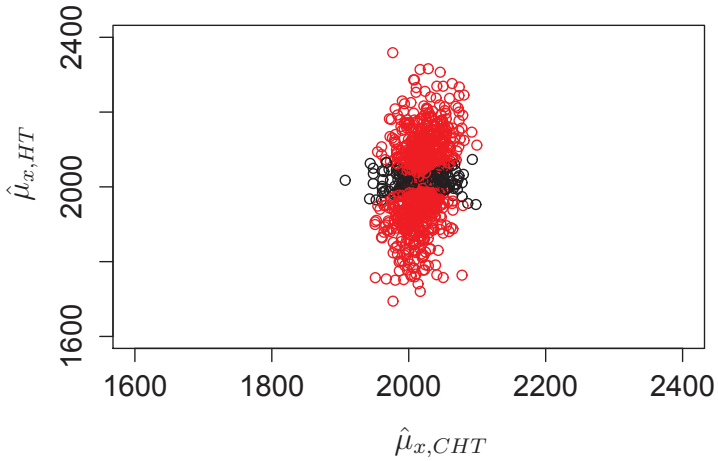


Figure 6.1: Comparison between $\hat{\mu}_{x,CHT}$ and $\hat{\mu}_{x,CH}$

The population size is $N = 500$, the variable y is a quantitative variable drawn from a uniform distribution over the interval $[0, 4000]$. The population is divided into 4 strata corresponding to the values of y_k (if $y_k \in [0, 1000[$ then k belongs to the strata 1 and so on ...). The auxiliary information will be the size of each strata in the population. In this example, we get $N_1 = 123$, $N_2 = 123$, $N_3 = 132$ and $N_4 = 122$.

The finite population stays fixed and we simulate with the software R $K = 10^3$ simple random samples of size $n = 100$. Two estimators of the mean $\mu_y = \frac{1}{N}\sum_{k\in U} y_k$ are computed and compared. The first one is the HT estimator: $\hat{\mu}_{y,HT} = \frac{1}{n}\sum_{k\in s} y_k$ and the second one is the conditional estimator: $\hat{\mu}_{y,CHT} = \frac{1}{N}\sum_h \sum_{k\in U_h} y_k \frac{N_h}{n_h(s)} I_{[k\in s]}$.

On Figure 6.1, we can see the values of $\hat{\mu}_{y,HT}$ and $\hat{\mu}_{y,CHT}$ for each of the 10^3 simulations. The red dots are those for which the conditional estimation is closer to the true value $\mu_y = 2019.01$ than the unconditional estimation; red dots represents 83.5% of the simulations. Moreover, the empirical variance of the conditional estimator is clearly smaller than the empirical variance of the unconditional estimator.

This is completely coherent with the results obtained for the post-stratification estimator in an model-assisted approach (see Särndal et Al.(1992) for example). However, what is new and fundamental in the conditional approach, is to understand that for one fixed sample, the conditional bias and variance are much more reliable than the unconditional bias and variance. The theoretical study of the conditional

variance estimation is a subject still to be developed.

6.3.4 Discussion

1. The traditional sampling strategy is defined as a couple (sampling design + estimator). We propose to define here the strategy as a triplet (sampling design + conditional sampling probability + estimator).
2. We have conditioned on the event: $A_0 = \{s \in \mathcal{S}, \forall h \in [1, H]\ n_h(s) = n_h(s_0)\}$. Under a SRS, it is similar to use the HT estimators of the sizes of the strata in the conditioning, that is to use $\mathbf{\Phi}(s) = (\hat{N}_1(s), ..., \hat{N}_H(s))^t$, where $\hat{N}_h(s) = \sum_{k \in U_h} \frac{I_{[k \in s]}}{\pi_k} = \frac{N}{n} n_h(s)$. Then, $A_0 = \{s \in \mathcal{S}, \mathbf{\Phi}(s) = \mathbf{\Phi}(s_0)\}$. We will see in Section 6.5 the importance of this remark.
3. The CHT estimations of the sizes of the strata are equal to the true strata sizes N_h, which means that the CHT estimations, in this setting, have the calibration property for the auxiliary information of the size of the strata. Hence, conditional inference gives a theoretical framework for the current practice of calibration on auxiliary variables.

6.4 A Posteriori Conditional Poisson Stratification

Rao(1985), Tillé(1999) and Andersson (2004) mentioned that a posteriori stratification in a more complex setting than an initial SRS is not a trivial task, and that one must rely on approximate procedures. In this section, we show that it is possible to determine the conditional sampling design and to compute exactly the conditional inclusion probabilities for an a posteriori stratification with a conditional Poisson sampling of size n.

6.4.1 Conditional Inference

Let $\tilde{p}(s) = \prod_{k \in s} p_k \prod_{k \in \bar{s}} (1 - p_k)$ be a Poisson sampling with inclusion probabilities $\mathbf{p} = (p_1, \ldots, p_N)^\top$, where $p_k \in]0, 1]$ and $\bar{s}$ is the complement of s in U. Under a Poisson sampling, the units are selected independently.
By means of rejective technics, a conditional Poisson sampling of size n can be implemented from the Poisson sampling. Then, the sampling design is:

$$p(s) = K^{-1} \mathbf{1}_{|s|=n} \prod_{k \in s} p_k \prod_{k \in \bar{s}} (1 - p_k),$$

where $K = \sum_{s, |s|=n} \prod_{k \in s} p_k \prod_{k \in \bar{s}} (1 - p_k)$.
The inclusion probabilities $\pi_k = f_k(U, \mathbf{p}, n)$ may be computed by means of a recursive method:

$$f_k(U, \mathbf{p}, n) = \frac{p_k}{1 - p_k} \frac{n}{\sum_{l \in U} \frac{p_l}{1 - p_l} (1 - f_l(U, \mathbf{p}, n - 1))} (1 - f_k(U, \mathbf{p}, n - 1))$$

where $f_k(U, \mathbf{p}, 0) = 0$.

This fact was proven by Chen et al.((Chen et al., 1994)) and one can also see Deville ((Deville, 2000)), Matei and Tillé ((Matei and Tillé, 2005)), and Bondesson(2010). An alternative proof is given in Annex 1.

It is possible that the initial π_k of the conditional Poisson sampling design are known instead of the p_k's. Chen et al.(1994) have shown that it is possible to inverse the functions $f_k(U, \mathbf{p}, n)$ by the means of an algorithm which is an application of the Newton method. One can see also Deville ((Deville, 2000)) who gave an enhanced algorithm.

Assume that a posteriori, thanks to some auxiliary information, the population is stratified in H strata U_h, $h \in [1, H]$. The size of the strata U_h is known to be equal to N_h, and the size of the sub-sample s_h into U_h is $n_h(s_0) > 0$. We consider the event $A_0 = \{s \in \mathcal{S}, \forall h \in [1, H], n_h(s) = n_h(s_0)\}$.

Proposition 8. *With an initial conditional Poisson sampling of size n:*

1. *The probability conditional to the sub-samples sizes of the "a posteriori strata", $p^{A_0}(s) = p(s/A_0)$, is the probability law of a **stratified sampling** with (independent) **conditional Poisson sampling of size** $n_h(s_o)$ **in each stratum**,*

2. *The conditional inclusion probability $\pi_k^{A_0}$ of an element k of the strata U_h is the inclusion probability of a conditional Poisson sampling of size $n_h(s_o)$ in a population of size N_h.*

Proof. 1. For a conditional Poisson of fixe size n, a vector $(p_1, \ldots, p_N)^\top$ exists, where $p_k \in]0, 1]$, such that:

$$p(s) = K^{-1} \mathbf{1}_{|s|=n} \prod_{k \in s} p_k \prod_{k \in \bar{s}} (1 - p_k),$$

where $K = \sum_{s, |s|=n} \prod_{k \in s} p_k \prod_{k \in \bar{s}} (1 - p_k)$.
We remind that $A_0 = \{s \in \mathcal{S}, \forall h \in [1, H], n_h(s) = n_h(s_0)\}$
Then:

$$\begin{aligned} p(A_0) &= K^{-1} \tilde{p} \left(\bigcap_{h \in [1,H]} [n_h(s) = n_h(s_0)] \right) \\ &= K^{-1} \prod_{h \in [1,H]} \tilde{p} \left([n_h(s) = n_h(s_0)], \right) \end{aligned}$$

where, $\tilde{p}(.)$ is the law of the original Poisson sampling. Let $s \in A_0$, then:

$$\begin{aligned} p^{A_0}(s) &= \frac{p(s)}{p(A_0)} \\ &= \frac{K^{-1}\prod_{h=1,\ldots,H}\left[\prod_{k\in s_h} p_k \prod_{k\in \bar{s}_h}(1-p_k)\right]}{K^{-1}\prod_{h\in[1,H]}\tilde{p}([n_h(s)=n_h(s_0)])} \\ &= \prod_{h=1,\ldots,H} \frac{\prod_{k\in s_h} p_k \prod_{k\in \bar{s}_h}(1-p_k)}{\tilde{p}([n_h(s)=n_h(s_0)])} \\ &= \prod_{h=1,\ldots,H} \frac{\prod_{k\in s_h} p_k \prod_{k\in \bar{s}_h}(1-p_k)}{\sum_{s_h,|s_h|=n_h(s_0)}\prod_{k\in s-h} p_k \prod_{k\in \bar{s}_h}(1-p_k)}, \end{aligned}$$

which is the sampling design of a stratified sampling with independent conditional Poisson sampling of size $n_h(s_o)$ in each stratum.
2. follows immediately. □

Definition 6. *In the context of conditional inference on the sub-sample sizes of posteriori strata, under an initial conditional Poisson sampling of size n, the* ***conditional HT estimator*** *of the total t_y is:*

$$\hat{t}_{y,CHT} = \sum_{k\in s} \frac{y_k}{\pi_k^{A_0}}.$$

The conditional variance can be estimated by means of one of the approximated variance formulae developed for the conditional Poisson sampling of size n. See for example Matei and Tillé(2005), or Andersson(2004).

6.4.2 Simulations

We take the same population as in subsection 6.3.3. The sampling design is now a conditional Poisson sampling of size $n = 100$. The probabilities p_k of the underlying Poisson design have been generated randomly, in order that $\sum_{k\in U} p_k = n$ and $p_k \in [0.13; 0.27]$.
$K = 10^3$ simulations were performed. Figure 6.2 shows that the punctual estimation of the mean of y is globally better for conditional inference. According to 77.3% of the simulations the conditional estimator is better than the unconditional estimator (red dots). The empirical variance as well is clearly better for the conditional estimator.

6.5 Conditioning on the Horwitz-Thompson estimator of an auxiliary variable

In the previous sections, we used the sub-sample sizes in the strata $n_h(s)$ to condition on. The good performances of this conditional approach result from the fact that the sizes of the sub-sample are important characteristics of the sample that are often used at the sampling stage. So, it was not surprising that the use of this

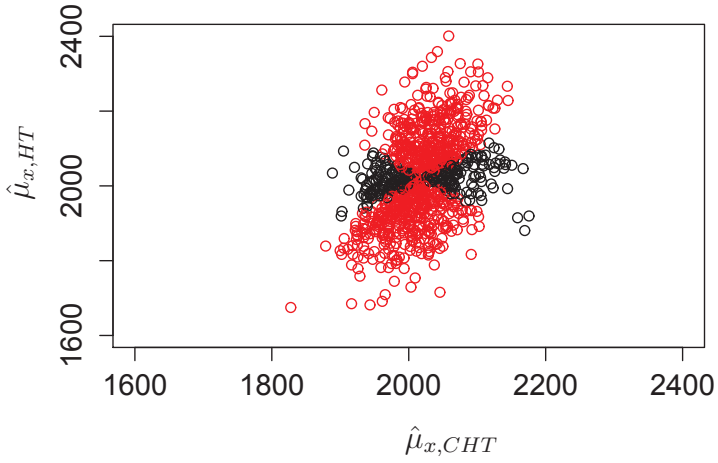


Figure 6.2: Comparison between $\hat{\mu}_{x,CHT}$ and $\hat{\mu}_{x,HT}$

information at the estimation stage would enhance the conditional estimators.

Another statistic that characterizes the representativeness of a sample is its HT estimator of the mean μ_x (or total t_x) of an auxiliary variable. This statistic is used at the sampling stage in balanced sampling for example. So, as the sub-sample sizes into the strata, this statistic should produce good results in a conditional approach restraining the inference to the samples for which the HT estimation of μ_x are equal to the value $\hat{\mu}_0 = \hat{\mu}_{x,HT}(s_0)$ of the selected sample s_0 .

In fact, we want the (conditional) set of the possible samples to be large enough in order that all conditional inclusion probabilities be different from zero. It is therefore convenient to consider the set of samples that give HT estimations not necessarily strictly equal to $\hat{\mu}_0$ but close to $\hat{\mu}_0$. Let $\varphi = [\hat{\mu}_0 - \varepsilon, \hat{\mu}_0 + \varepsilon]$, for some arbitrary quantity $\varepsilon > 0$. This idea is similar to the tolerance in the rejectif sampling of Fuller (2009)

The set A_φ of possible samples in our conditional approach will be:

$$A_\varphi = \{s \in \mathcal{S},\ \hat{\mu}_{x,HT}(s) \in [\widehat{\mu}_0 - \varepsilon, \hat{\mu}_0 + \varepsilon]\}.$$

The conditional inclusion probability of a unit k is:

$$\begin{aligned} \pi_k^{A_\varphi} &= p\left([k \in s] / \ [\hat{\mu}_{x,HT}(s) \in [\hat{\mu}_0 - \varepsilon, \hat{\mu}_0 + \varepsilon]]\right) \\ &= \frac{p\left(\{s \in \mathcal{S},\ k \in s \text{ and } \hat{\mu}_{x,HT}(s) \in [\hat{\mu}_0 - \varepsilon, \hat{\mu}_0 + \varepsilon]\}\right)}{p(A_\varphi)}. \end{aligned}$$

If $\hat{\mu}_0 = \mu_X$ then we are in a good configuration, because we are in a balanced sampling situation and the $\pi_k^{A_\varphi}$ will certainly stay close to the π_k.

If $\widehat{\mu}_0 \gg \mu_X$ say, then the sample s_0 is unbalanced, which means that in average, its units have a too large contribution x_k/π_k, either because they are too big (x_k large) or too heavy ($d_k = \frac{1}{\pi_k}$ too large). In this case, the samples in A_φ are also ill-balanced, because balanced on $\widehat{\mu}_0$ instead of μ_X: $\sum_{k\in s} \frac{x_k}{\pi_k} \approx \widehat{\mu}_0$. But conditioning on this information will improve the estimation. Indeed, the $\pi_k^{A_\varphi}$ will be different from the π_k. For example, a unit k with a big contribution ($\frac{x_k}{\pi_k}$ large) has more chance to be in a sample of A_φ than a unit l with a small contribution. So, we can expect that $\pi_k^{A_\varphi} > \pi_k$ and $\pi_l^{A_\varphi} < \pi_l$. And, in consequence, the conditional weight d_k^φ will be lower than d_k and d_l^φ higher than d_l, which will "balance" the samples of A_φ.

Discussion:

– we can use different ways in order to define the subset φ. One way is to use the distribution function of $\Phi(s)$, denoted $G(u)$ and to define φ as a symmetric interval:

$$\varphi = \left[G^{-1}(\max\{G(\Phi(s_0)) - \frac{\alpha}{2}, 0\}), G^{-1}(\min\{G(\Phi(s_0)) + \frac{\alpha}{2}, 1)\}\right],$$

where $\alpha = 5\%$ for example.

Hence,

$$A_\varphi = \{s \in \mathcal{S}, \\ \Phi(s) \in \left[G^{-1}(\max\{G(\Phi(s_0)) - \frac{\alpha}{2}, 0\}), G^{-1}(\min\{G(\Phi(s_0)) + \frac{\alpha}{2}, 1\})\right]\},$$

and $p(A_\varphi) \leq \alpha$.

As the *cdf* $G(u)$ is unknown in general, one has to replace it by an estimated *cdf* of $\Phi(s)$, denoted $\hat{G}_K(u)$, computed by means of simulations.

6.6 Generalization: Conditional Inference Based on Monte Carlo simulations.

In this section, we consider a general initial sample design $p(s)$ with the inclusion probabilities π_k. We condition on the event $A_\varphi = \mathbf{\Phi}^{-1}(\varphi) = \{s \in \mathcal{S}, \mathbf{\Phi}(s) \in \varphi\}$. For example, we can use $\Phi(s) = \sum_{k\in s} \frac{\mathbf{x}_k}{\pi_k}$ the unconditional HT estimator of t_x and $\varphi = [\varphi_1, \varphi_2]$ an interval that contains $\Phi(s_0) = \sum_{k\in s_0} \frac{\mathbf{x}_k}{\pi_k}$, the HT estimation of t_x with the selected sample s_0. In other words, we will take into account the information that the HT estimator of the total of the auxiliary variable x lies in some region φ.

The mathematical expression of $\pi_k^{A_\varphi}$ is straightforward:

$$\pi_k^{A_\varphi} = p([k \in s]/A_\varphi) = \frac{\sum_s p(s)\mathbf{1}_{s\in A_\varphi}\mathbf{1}_{[k\in s]}}{p(A_\varphi)}.$$

But effective computation of the $\pi_k^{A_\varphi}$'s may be not trivial if the distribution of $\mathbf{\Phi}$ is complex. Tillé((Tillé, 1998)) used an asymptotical approach to solve this problem when $\Phi(s) = \sum_{k \in s} \frac{x_k}{\pi_k} \mathbf{1}_{[k \in s]}$; he has used normal approximations for the conditional and unconditional laws of Φ.
In the previous sections, we have given examples where we were able to compute the $\pi_k^{A_\varphi}$'s (and actually the $p^{A_\varphi}(s)$'s) exactly. In this section, we give a general Monte Carlo method to compute the $\pi_k^{A_\varphi}$.

6.6.1 Monte Carlo

We will use Monte Carlo simulations to estimate $\mathbb{E}(\mathbf{1}_{A_\varphi} \mathbf{1}_{[k \in s]})$ and $\mathbb{E}(\mathbf{1}_{A_\varphi})$. We repeat independently K times the sample selection with the sampling design $p(s)$, thus obtaining a set of samples $(s_1, \ldots, s_K)$. For each simulation i, we compute $\mathbf{\Phi}(s_i)$ and $I_{A_\varphi}(s_i)$. Then we compute $N + 1$ statistics:

$$\begin{aligned} M^{A_\varphi} &= \sum_{i=1}^{K} \mathbf{1}_{A_\varphi}(s_i) \\ \forall k \in U, M_k^\varphi &= \sum_{i=1}^{K} \mathbf{1}_{A_\varphi}(s_i) \mathbf{1}_{[k \in s_i]} \end{aligned}$$

We obtain a consistent estimator of $\pi_k^{A_\varphi}$, as $K \to +\infty$:

$$\hat{\pi}_k^{A_\varphi} = \frac{M_k^\varphi / K}{M^{A_\varphi} / K} = \frac{M_k^\varphi}{M^{A_\varphi}} \tag{6.1}$$

6.6.2 Point and variance estimations in conditional inference

Definition 7. *The Monte Carlo estimator of the total t_y is the conditional Horvitz-Thompson estimator of t_y after replacing the conditional inclusion probabilities by their Monte Carlo approximations:*

$$\hat{t}_{y,MC} = \sum_{k \in s_0} \frac{1}{\hat{\pi}_k^{A_\varphi}} y_k$$

The Monte Carlo estimator of the variance of $\hat{t}_{y,MC}$ is:

$$\widehat{\mathbb{V}}(\hat{t}_{y,MC}) = \sum_{k,l \in s_0} \frac{1}{\hat{\pi}_{k,l}^{A_\varphi}} \frac{y_k}{\hat{\pi}_k^{A_\varphi}} \frac{y_l}{\hat{\pi}_l^{A_\varphi}} (\hat{\pi}_{k,l}^{A_\varphi} - \hat{\pi}_k^{A_\varphi} \hat{\pi}_l^{A_\varphi}),$$

where

$$\hat{\pi}_{k,l}^{A_\varphi} = \frac{\sum_{i=1}^{K} t_y \mathbf{1}_{A_\varphi}(s_i) \mathbf{1}_{[k \in s_i]} \mathbf{1}_{[l \in s_i]}}{\sum_{i=1}^{K} \mathbf{1}_{A_\varphi}(s_i)}.$$

Fattorini((Fattorini, 2006)) established that $\hat{t}_{y,MC}$ is asymptotically unbiased as $M^{A_\varphi} \to \infty$, and that its mean squared error converges to the variance of $\hat{t}_{y,HT}$.

Thompson and Wu ((Thompson Mary and Wu, 2008)) studied the rate of convergence of the estimators $\hat{\pi}_k^{A_\varphi}$ and of the estimator $\hat{t}_{y,MC}$ following Chebychev's inequality. Using normal approximation instead of the Chebychev's inequality gives more precise confidence intervals. We have thus a new confidence interval for $\hat{\pi}_k^{A_\varphi}$:

$$p\left(|\hat{\pi}_k^{A_\varphi} - \pi_k^{A_\varphi}| < F^{-1}((1-\alpha)/2)\sqrt{\frac{1}{4M^{A_\varphi}}}\right) \leq \alpha,$$

where F is the distribution function of the normal law $\mathcal{N}(0,1)$.
As for the relative bias, standard computation leads to:

$$\begin{aligned} p\left(\frac{|\hat{t}_{y,CHT} - \tilde{t}_{y,CHT}|}{\hat{t}_{y,CHT}} \leq \varepsilon\right) &\geq 1 - 4 \times \sum_{k \in s}\left[1 - F\left(\frac{\epsilon}{1+\epsilon}\sqrt{M^{A_\varphi}\pi_k^{A_\varphi}}\right)\right] \\ &\geq 1 - 4n\frac{1}{\sqrt{2\pi}}\frac{1+\epsilon}{\sqrt{M^{A_\varphi}\epsilon^2\pi_0}} \cdot e^{-\left(\frac{M^{A_\varphi}\epsilon^2}{(1+\epsilon)^2}\pi_0\right)}, \end{aligned} \quad (6.2)$$

where $\pi_0 = \min\{\pi_k^{A_\varphi}, k \in U\}$. We used the inequality $1 - F(u) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-u^2}}{u}$ which is verified for large u.

The number K of simulations is set so that $\sum_{i=1}^{K} I_{A_\varphi}(s_i)$ reaches a pre-established M^{A_φ} value. Because of our conditional framework, K is a stochastic variable which follows a negative binomial distribution and we have $E(K) = \dfrac{M^{A_\varphi}}{p(A_\varphi)}$. For instance, if $p(A_\varphi) = 0.05 = 5\%$, with $M^{A_\varphi} = 10^6$, we expect $E(K) = 2.10^7$ simulations.

6.7 Conditional Inference Based on Monte Carlo Method in Order to Adjust for Outlier and Strata Jumper

We will apply the above ideas to two examples close to situations that can be found in establishments surveys: outlier and strata jumper.
We consider an establishments survey, performed in year "*t+1*", and addressing year "t". The auxiliary information x which is the turnover of the year "t" is not known at the sampling stage but is known at the estimation stage (this information may come from, say, the fiscal administration).

6.7.1 Outlier

In this section, the auxiliary variable x is simulated following a gaussian law, more precisely $x_k \sim \mathfrak{N}(8\ 000, (2\ 000)^2)$ excepted for unit $k = 1$ for which we assume that $x_1 = 50\ 000$. The unit $k = 1$ is an outlier. The variable of interest y is simulated by the linear model

$$y_k \ = \ 1000 \ + \ 0.2\ x_k \ + \ u_k,$$

where $u_k \sim \mathfrak{N}(0, (500)^2)$, u_k is independent from x_k. The outcomes are $\mu_x = 8\ 531$ and $\mu_y = 2\ 695$.

We assume that the sampling design of the establishments survey is a SRS of size

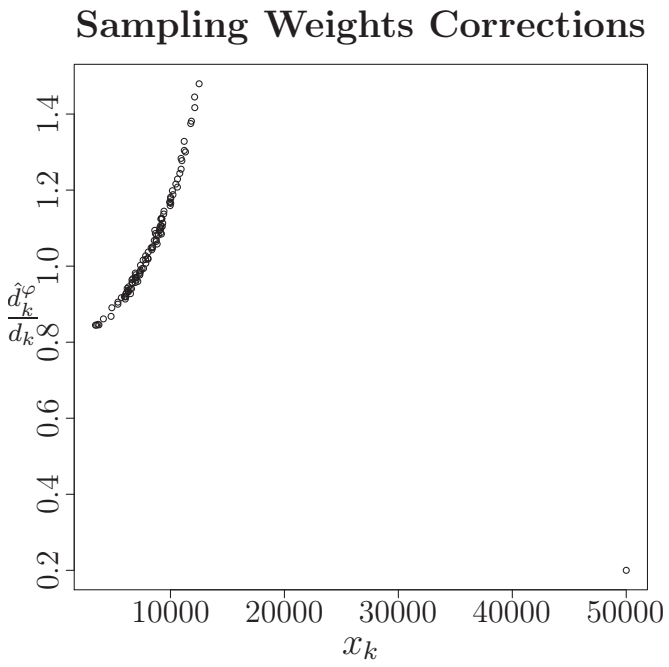


Figure 6.3: Outlier, sampling weight corrections

$n = 20$ out of the population U of size $N = 100$ and that the selected sample s_0 contains the unit $k = 1$. For this example, we have repeated the sample selection until the unit 1 has been selected in s_0.

We obtain $\Phi(s_0) = \hat{\mu}_{x,HT}(s_0) = 9\ 970$, which is 17% over the true value $\mu_x = 8\ 531$ and $\hat{\mu}_{y,HT}(s_0) = 3\ 039$ (recall that the true value of μ_y is 2 695).

We set Φ and φ as in section 6.5 and we use Monte Carlo simulations in order to compute the conditional inclusion probabilities $\hat{\pi}_k^{A_\varphi}$. Each simulation is a selection of a sample following a SRS of size $n = 20$ from the fixed population U. Recall that the value of x_k will eventually be known for any unit $k \in U$.

Actually, we use two sets of simulations. The first set is performed in order to estimate the *cdf* of the statistic $\Phi(s) = \hat{\mu}_{x,HT}(s)$ which will be used to condition on. This estimated *cdf* will enable us to construct the interval φ. More precisely, we choose the interval $\varphi = [9\ 793, 10\ 110]$ by the means of the estimated *cdf* of $\Phi(s) = \hat{\mu}_{y,HT}(s)$ and so that $p\left([\hat{\mu}_{x,HT}(s) \in [9\ 793, \hat{\mu}_{x,HT}(s_0)]]\right) = \frac{\alpha}{2} = 2.5\% = p\left([\hat{\mu}_{x,HT}(s) \in [\hat{\mu}_{x,HT}(s_0), 10\ 110]]\right)$.

A_φ is then the set of the possible samples in our conditional approach:

$$A_\varphi = \{s \in \mathcal{S},\ \hat{\mu}_{x,HT}(s) \in [9\ 793, 10\ 110]\}.$$

Note that $p\left([\hat{\mu}_{x,HT}(s) \in [9\ 793, 10\ 110]]\right) = \alpha = 5\%$. A_φ typically contains samples

that over-estimate the mean of x.

The second set of Monte Carlo simulations consists in $K = 10^6$ sample selections with a SRS of size $n = 20$ performed in order to estimate the conditional inclusion probabilities $\hat{\pi}_k^{A_\varphi}$. 49 782 (4.98%) simulated samples fall in A_φ, and among them, 49 767 samples contain the outlier, which correspond to the estimated conditional inclusion probability of the outlier: $\hat{\pi}_1^{A_\varphi} = 0.9997$. It means that almost all the samples of A_φ contain the outlier that is mainly responsible for the over-estimation because of its large value of the variable x!
The weight of the unit 1 has changed a lot, it has decreased from $d_k = \frac{1}{0.2} = 5$ to $\hat{d}_k^{A_\varphi} = 1.0003$. The conditional sampling weights of the other units of s_0 are more comparable to their initial weights $d_k = 5$ (see Figure 6.3).

The conditional MC estimator $\hat{\mu}_{y,MC}(s) = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\hat{\pi}_1^{A_\varphi}}$ leads to a much better estimation of μ_y: $\hat{\mu}_{y,MC}(s_0) = 2\ 671$.

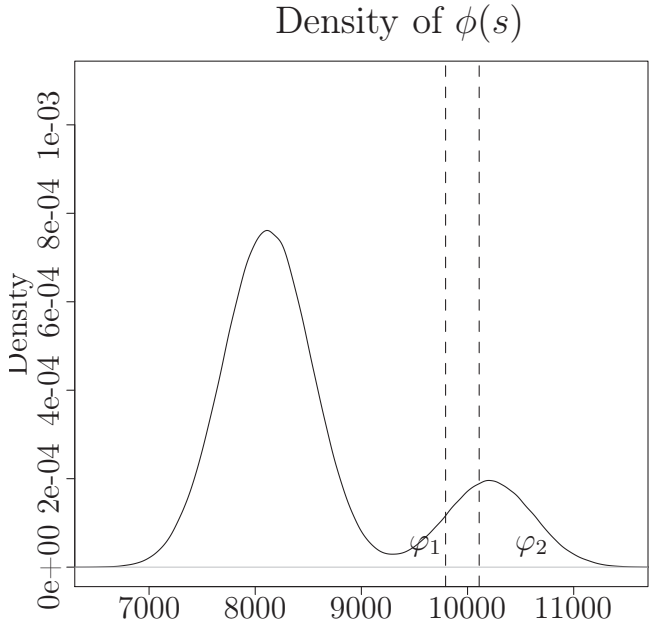


Figure 6.4: Outlier, Density of $\Phi(s) = \hat{\mu}_{x,HT}(s)$

Figure 6.3 gives an idea of the conditional inclusion probabilities for all the units of U. Moreover, this graph shows that the correction of the sampling weights $\frac{\hat{d}_k^{A_\varphi}}{d_k} = \frac{\pi_k}{\hat{\pi}_k^{A_\varphi}}$ is not a monotonic function of x_k, which is in big contrast with calibration techniques which only uses monotonic functions for weight correction purposes.

A last remark concerns the distribution of the statistics $\Phi(s) = \hat{\mu}_{x,HT}(s)$. Figure

6.4 shows an unconditional distribution with 2 modes and far from gaussian. This shows that in presence of outlier, we can not use the method of Tillé ((Tillé, 1999)), which assumes a normal distribution for $\hat{\mu}_{x,HT}(s)$.

6.7.2 Strata Jumper

In this section, the population U is divided into 2 sub-populations: the small firms and the large firms. Let us say that the size is appreciated thanks to the turnover of the firm. Official statistics have to be disseminated for this 2 different sub-populations. Hence, the survey statistician has to split the population into 2 strata corresponding to the sub-populations. This may not be an easy job because the size of firms can evolve from one year to another.
Here we assume that, at the time when the sample is selected, the statistician does not know yet the auxiliary information x of the turnover of the firm for the year "n", more precisely the strata the firm belongs to for the year "n". Let us assume that he only knows this information for the previous year,"*n-1*". This information is denoted by z. In practice, small firms are very numerous and the sampling rate for this strata is chosen low. On the contrary, large firms are less numerous and their sampling rate is high.

When a unit is selected among the small firms but eventually happens to be a large unit of year "n", we call it a strata jumper. At the estimation stage, when the information x becomes available, this unit will obviously be transferred to strata 2 . This will bring a problem, not due to its y-value (which may well be typical in strata 2) but to its sampling weight, computed according to strata 1 (the small firms), and which will appear to be very large in comparison to the other units in strata 2 at the estimation stage.

In our simulations, the population U is split in 2 strata, by means of the auxiliary variable z: U_1^z, of size $N_1^z = 10\ 000$, is the strata of presumed small firms and U_2^z, of size $N_2^z = 100$, the strata of presumed large firms.
The auxiliary variable x, which is the turnover of the year "n" known after collection, is simulated under a gaussian law $\mathfrak{N}(8\ 000, (2\ 000)^2)$ for the units of the strata U_2^z and for one selected unit of the strata U_1^z. Let us say that this unit, the strata jumper, is unit 1.
Our simulation gives $x_1 = 8\ 002$. The variable of interest y is simulated by the linear model $y_k = 1000 + 0.2\, x_k + u_k$, where $u_k \sim \mathfrak{N}(0, (500)^2)$, u_k and x_k independent. We do not simulate the value of x and y for the other units of the strata U_1^z because we will focus on the estimation of the mean of y for the sub-population of large firms of year n U_2^x: $\mu_{y,2} = \frac{1}{N_2} \sum_{k \in U_2^x} y_k$. We find $\mu_{x,2} = 8\ 138$ and $\mu_{y,2}$ is 2 606.

The sampling design of the establishments survey is a stratified SRS of size $n_1 = 400$ in U_1^z and $n_2 = 20$ in U_2^z. We assume that the selected sample s_0 contains the unit $k = 1$. In practice, we repeat the sample selection until the unit 1 (the strata jumper) has been selected.

As previously, Φ and φ are defined as in Section 6.5.
We use Monte Carlo simulations in order to compute the conditional inclusion probabilities $\hat{\pi}_k^{A_\varphi}$. A simulation is a selection of a sample with stratified SRS of size $n_1 = 400$ in U_1^z and $n_2 = 20$ in U_2^z.

We choose the statistic $\Phi(s) = \hat{\mu}_{x,2,HT}(s)$ in order to condition on. $K = 10^6$ simulations are performed in order to estimate the *cdf* of $\Phi(s)$ and the conditional inclusion probabilities.
Our simulations give $\Phi(s_0) = \hat{\mu}_{x,2,HT}(s_0) = 9\ 510$, which is far from the true value $\mu_{x,2} = 8\ 138$ and $\hat{\mu}_{y,2,HT}(s_0) = 3\ 357$ (recall that the true value of $\mu_{y,2}$ is 2 606). We choose the interval $\varphi = [8\ 961, 10\ 342]$ by the means of the estimated *cdf* of $\Phi(s) = \hat{\mu}_{x,2,HT}(s)$ and so that $p\left([\hat{\mu}_{x,2,HT}(s) \in [8\ 961, 10\ 342]]\right) = \alpha = 5\%$.
A_φ is then the set of the possible samples in our conditional approach:

$$A_\varphi = \{s \in \mathcal{S},\ \hat{\mu}_{x,HT}(s) \in [8\ 961, 10\ 342]\}.$$

All samples in A_φ over-estimate the mean of x.

Among the 10^6 simulations, 49 778 simulated samples (4.98%) belongs to A_φ. 55% of them contains the strata jumper, which gives the estimated conditional inclusion probability of the strata jumper $\hat{\pi}_1^{A_\varphi} = 0.55$. It is not a surprise that the strata jumper is in one sample of A_φ over two. Indeed, its initial sampling weight $d_1 = \frac{10\ 000}{400} = 25$ is high in comparison to the weights $d_k = \frac{100}{20} = 5$ of the other selected units of the strata U_2^x, and its contribution $\frac{25x_1}{N_2}$ contributes to over-estimate the mean of x .

The conditional inclusion probabilities for the other units of U_2^x are comparable to their initial $\pi_k = 0.2$ (see Figure 6.5).

The conditional MC estimator $\hat{\mu}_{y,2,MC}(s) = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\hat{\pi}_1^{A_\varphi}}$ leads to a better estimation of $\mu_{y,2}$: $\hat{\mu}_{y,2,MC}(s_0) = 2\ 649$.

Figure 6.5 shows that sampling weights correction is here a non-monotonic function of the variable x. We point out that the usual calibration method would not be able to perform this kind of weights correction because the calibration function used to correct the weights should be monotonic.
Similarly to the outlier setting, the unconditional distribution of the statistics $\Phi(s) = \hat{\mu}_{x,2,HT}(s)$ has 2 modes and is far from gaussian.

6.8 Conclusion

At the estimation stage, a new auxiliary information can reveal that the selected sample is imbalanced. We have shown that a conditional inference approach can take into account this information and leads to a more precise estimator than the

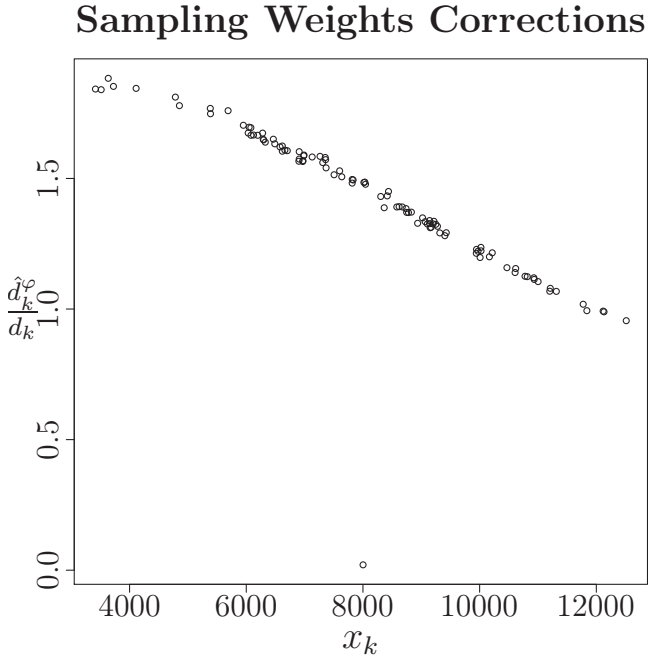


Figure 6.5: Strata Jumper, Sampling Weight Corrections

unconditional Horvitz-Thompson estimator in the sense that the conditional estimator is unbiased (conditionally and unconditionally) and that the conditional variance is more rigorous in order to estimate the precision a posteriori.
In practise, we recommend to use Monte Carlo simulations in order to estimate the conditional inclusion probabilities.
This technic seems particularly adapted to the treatment of outliers and strata-jumpers.

6.9 Annex 1: Inclusion Probability with Conditional Poisson Sampling

Proof. The event $\left[\sum_{l\in U, l\neq k} I_{[l\in s]} = n-1\right]$ is independent of the events $[I_{[k\in s]}=0]$ and $[I_{[k\in s]}=1]$ in the Poisson model. So we can write:

$$\begin{aligned} p\left(\left[\sum_{l\in U, l\neq k} I_{[l\in s]} = n-1\right]\right) &= p\left(\left[\sum_{l\in U, l\neq k} I_{[l\in s]} = n-1\right]/[I_{[k\in s]}=0]\right) \quad (6.3)\\ &= p\left(\left[\sum_{l\in U, l\neq k} I_{[l\in s]} = n-1\right]/[I_{[k\in s]}=1]\right) \quad (6.4)\end{aligned}$$

Equation (6.3) gives:

$$\begin{aligned} &p\left(\left[\sum_{l\in U, l\neq k} I_{[l\in s]} = n-1\right]/[I_{[k\in s]}=0]\right)\\ &\quad = p\left(\left[\sum_{l\in U} I_{[l\in s]} = n-1\right]/[I_{[k\in s]}=0]\right)\\ &\quad = \frac{p\left(\left[\sum_{l\in U} I_{[l\in s]} = n-1\right]\right) p\left([I_{[k\in s]}=0]/\left[\sum_{l\in U} I_{[l\in s]} = n-1\right]\right)}{p\left([I_{[k\in s]}=0]\right)}\\ &\quad = \frac{p\left(\left[\sum_{l\in U} I_{[l\in s]} = n-1\right]\right)(1-f_k(N,\mathbf{p},n-1))}{1-p_k}, \end{aligned}$$

and equation (6.4) gives:

$$\begin{aligned} &p\left(\left[\sum_{l\in U, l\neq k} I_{[l\in s]} = n-1\right]/[I_{[k\in s]}=1]\right)\\ &\quad = p\left(\left[\sum_{l\in U} I_{[l\in s]} = n\right]/[I_{[k\in s]}=1]\right)\\ &\quad = \frac{p\left(\left[\sum_{l\in U} I_{[l\in s]} = n\right]\right) p\left([I_{[k\in s]}=1]/\left[\sum_{l\in U} I_{[l\in s]} = n\right]\right)}{p\left([I_{[k\in s]}=1]\right)}\\ &\quad = \frac{p\left(\left[\sum_{l\in U} I_{[l\in s]} = n\right]\right) f_k(N,\mathbf{p},n)}{p_k}. \end{aligned}$$

So we have:

$$\begin{aligned} f_k(U,\mathbf{p},n) &= (1-f_k(U,\mathbf{p},n-1))\frac{p_k}{1-p_k}\frac{p\left(\left[\sum_{l\in U} I_{[l\in s]} = n-1\right]\right)}{p\left(\left[\sum_{l\in U} I_{[l\in s]} = n\right]\right)}\\ &= (1-f_k(U,\mathbf{p},n-1))\frac{p_k}{1-p_k}h(U,\mathbf{p},n) \end{aligned}$$

And we can use the property $\sum_{k\in U} f_k(U,\mathbf{p},n) = \sum_{k\in U}\pi_k = n$ to compute $h(U,\mathbf{p},n)$ and conclude. □

Bibliography

Andersson, P. G. (2006). A conditional perspective of weighted variance estimation of the optimal regression estimator. *Journal of Statistical Planning and Inference*, 136(1):221–234.

Chen, X.-H., Dempster, A. P., and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3):457–469.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, France. In French.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2):269–278.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4):933–944.

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4):543–570.

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11:15–31.

Robinson, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82(399):826–831.

Särndal, C.-E., Swensson, B., and Wretman., J. (1992). *Model Assisted Survey Sampling*. New-Yok: Springer-Verlag.

Thompson Mary, E. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34(1):3–10.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *International Statistical Review*, 66:303–322.

Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities: comlex design. *Survey Methodology*, 25(1):57–66.

Chapter 7

Some aspects of balanced sampling

7.1 Introduction

Balanced sampling has received some attention in recent years; e.g., Deville et Tillé (2004), Chauvet and Tillé (2006), Fuller (2009b) and Legg and Yu (2010). Consider a finite population P of size N. We are interested in estimating the population total, $t_y = \sum_{i \in P} y_i$, where y denotes a characteristic of interest. Prior to sampling, we assume that a vector of auxiliary variables, $\mathbf{z}$ is available for all $i \in P$. The $\mathbf{z}$-variables are known as the designs variables. Let $\mathbf{Z}$ be the matrix of design information, whose i-th row is $\mathbf{z}_i^\top$. A sample $s \subset P$ is said to be $\psi\mathbf{z}$-balanced if

$$\hat{\mathbf{t}}_{\mathbf{z}}^{\psi} \equiv \sum_{i \in s} \psi_i^{-1} \mathbf{z}_i = \sum_{i \in P} \mathbf{z}_i \equiv \mathbf{t}_{\mathbf{z}}, \tag{7.1}$$

where $0 < \psi_i < 1$ for all $i \in P$. A sampling design satisfying (7.1) for all s is called a balanced sampling design. There exists a number of procedures leading to a balanced or approximately balanced sample, including the Cube method (Deville and Tillé, 2004) and rejective sampling (Fuller, 2009b).

Let π_i denote the inclusion probability attached to unit i with respect to the sampling design utilized to select the sample. In the case of the Cube method, the inclusion probabilities, $\pi_1, \ldots, \pi_N$ are fixed prior to sampling and the sample is selected so that the π_i's are exactly satisfied. However, there may not exist a sample s that satisfies (7.1) exactly and so we must accept that these balancing constraints be satisfied approximately.

In rejective sampling, we select repeated samples according to a basic sampling procedure until a selected sample meets a specified balancing tolerance. Commonly used basic sampling procedures include simple random sampling without replacement, stratified sampling and poisson sampling. Let p_i be inclusion probability attached to unit i with respect to the basic procedure. The p_i's are not to be confused with the π_i's which, for the rejective method, are the inclusion probabilities in the rejective sample (i.e., the sample selected at the end of the rejective process). Unlike for the Cube method, the π_i's are generally unknown for rejective sampling: only the p_i's are available prior to sampling. Let $\gamma > 0$ be a balancing tolerance

specified by the survey statistician. In rejective sampling (Fuller, 2009b), samples are rejected until

$$(\hat{\mathbf{t}}_{\mathbf{z}}^{p} - \mathbf{t}_{\mathbf{z}})^{\top} V_b(\hat{\mathbf{t}}_{\mathbf{z}}^{p})^{-1}(\hat{\mathbf{t}}_{\mathbf{z}}^{p} - \mathbf{t}_{\mathbf{z}}) \leq \gamma, \tag{7.2}$$

where $\hat{\mathbf{t}}_{\mathbf{z}}^{p} = \sum_{i \in s_b} p_i^{-1}\mathbf{z}_i$, s_b denotes a sample selected according to the basic procedure and $V_b(.)$ denotes the variance with respect to the basic procedure. Once again a sample selected according to the basic procedure s_b is not to be confused with the sample s selected at the end of the rejective process. In our notation, we have $p_i = P(i \in s_b)$ and $\pi_i = P(i \in s)$.

An important distinction between the Cube method and the rejective method is that, in the first, the inclusion probabilities π_i are exactly satisfied but one has not control on the (possible) discrepancy between estimates of the $\mathbf{z}$-variables and their true population total, whereas in the second, the discrepancy is perfectly controlled (this corresponds to the balancing tolerance γ) but the the inclusion probabilities π_i are unknown.

In this paper, we examine the properties (bias and variance) of estimation procedures with respect to the Cube method and rejective sampling. We adopt a design-based point of view. At this point, it is useful to recall some of the basic inferential principles in survey sampling. Let $\mathbf{I} = (I_1, \ldots, I_N)$, be the vector of sample selection indicators such that $I_i = 1$ if unit i is selected in the sample and $I_i = 0$, otherwise, and let $\mathbf{y} = (y_1, \ldots, y_N)$. We distinguish between two different distributions: (i) the distribution of $\mathbf{I}$, often called the sampling design, which we denote by $F(\mathbf{I}|\mathbf{Z})$. We assume that, given $\mathbf{Z}$, the distribution of $\mathbf{I}$ is completely specified. That is, we assume that there remains no residual relationship between $\mathbf{y}$ and $\mathbf{I}$ after accounting for $\mathbf{Z}$. This assumption is referred to as non-informative sampling; e.g., Pfeffermann (1993). (ii) The distribution of $\mathbf{y}$, which we denote by $F(\mathbf{y}|\mathbf{U})$, where $\mathbf{U}$ is a matrix of auxiliary information that may include or not the design variables $\mathbf{z}$. Based on these two distributions, there are several possible approaches to inference in survey sampling including (i) the design-based approach and (ii) the model-based approach.

In the design-based approach, properties of estimators are evaluated with respect to the sampling design. In other words, the vector $\mathbf{y}$ is held fixed and the only remaining source of randomness is the vector of sample indicators $\mathbf{I}$. The population total t_y is a fixed quantity that we wish to estimate. For conducting an inference with respect to the sampling design, the sample must be randomly selected from the population. In other words, design-based inference is not possible if the sample is selected in a non random fashion. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^{\top} = E(\mathbf{I}|\mathbf{Z})$ be the vector of first-order inclusion probabilities in the sample, which is also the first moment of the distribution, $F(\mathbf{I}|\mathbf{Z})$. Typically, samples are selected using a sampling procedure that ensures that, the inclusion probabilities $\pi_1, \ldots, \pi_N$ defined prior to sampling, are exactly satisfied. When the π_i's are known and are incorporated in the estimation procedures, the resulting estimators have good design properties such as (approximately) design-unbiasedness and design-consistency. This property is true irrespective of the form of $F(\mathbf{y}|\mathbf{Z})$ provided that some moment conditions are

satisfied; e.g., Isaki and Fuller (1982).
In contrast, the model-based approach does not require the sample to be selected in a random fashion. Properties of estimators are evaluated with respect to an assumed model for $F(\mathbf{y}|\mathbf{Z})$. The vector of sample selection indicators, $\mathbf{I}$, is held fixed. The population total t_y is now a random variable, which we would like to predict. We seek predictors of t_y that have good properties (bias, variance) under the assumed model. Royall (1976) derived the Best Linear Unbiased Predictor (BLUP) of t_y and showed that it has good properties (e.g., model unbiasdness and model-consistency) provided that the underlying model holds. In the model-based approach, complete reliance is placed on the model for $F(\mathbf{y}|\mathbf{Z})$.

At the estimation stage, suppose that a vector of calibration variables $\mathbf{u} = (u_1, \ldots, u_q)$ is available for all the sample units $i \in P$ and that the vector of population totals $t_{\mathbf{u}} = \sum_{i \in U} \mathbf{u}_i$ is known. Recall that the vector $\mathbf{u}$ may include or not the design variables $\mathbf{z}$. Let $\mathbf{U}$ be the matrix of calibration variables, whose i-th row is equal to $\mathbf{u}_i$. As an estimator of t_y, we consider a linear estimator of the form $\hat{t}_y = \sum_{i \in s} w_i y_i$, where w_i is weight attached to unit i. In the design-based literature, commonly used weights w_i include: $w_i = \pi_i^{-1}$ and

$$w_i = \pi_i^{-1} \left\{ 1 + \left(t_{\mathbf{u}} - \hat{t}_{\mathbf{u}} \right)^\top \left(\sum_{i \in s} \pi_i^{-1} \mathbf{u}_i \mathbf{u}_i^\top \right)^{-1} \mathbf{u}_i \right\}, \tag{7.3}$$

where $\hat{t}_{\mathbf{u}} = \sum_{i \in s} \pi_i^{-1} \mathbf{u}_i$. The choice $w_i = \pi_i^{-1}$ leads to the customary Horvitz-Thompson estimator, which is design-unbiased and design-consistent for t_y regardless of the y variable being estimated provided that the inclusion probabilities π_i's are known without error; e.g., Isaki and Fuller (1982). The choice (7.3) leads to the Generalized Regression (GREG) estimator, which is approximately design-unbiased and design-consistent for t_y regardless of the y variable being estimated; e.g., Wright (1983) and Särndal and Wright (1984).
When the π_i's are unknown (as it is the case for rejective sampling), we must settle for some approximation $\hat{\pi}_i$, say. If $\hat{\pi}_i$ provides a good approximation of π_i, we expect a Horvitz-Thompson type estimator, based on $\hat{\pi}_i$, to exhibit a small bias. However, this type of estimator may suffer from significant bias if $\hat{\pi}_i$ is a poor approximation of π_i. This is discussed in relation to rejective sampling in Section 4. For the GREG estimator, it is useful to introduce the following superpopulation model:

$$\begin{aligned} y_i &= \mathbf{u}_i^\top \boldsymbol{\beta} + \epsilon_i, \\ E_m(\epsilon_i|\mathbf{U}) = 0, V_m(\epsilon_i|\mathbf{U}) &= \sigma^2, \end{aligned} \tag{7.4}$$

where $\boldsymbol{\beta}$ and σ^2 are unknown parameters. When the true π_i's (if known) are incorporated in the GREG procedure, the resulting estimator is approximately design-unbiased and design consistent irrespective of whether or not model (7.4) holds. On the other hand, it is model-unbiased and model consistent for t_y if (7.4) holds, even if the π_i's are unknown and poorly estimated. More specifically, suppose that the π_i's are unknown and are replaced by some approximation $\hat{\pi}_i$ in (7.3). The resulting estimator is model-unbiased and model-consistent for t_y, even if the $\hat{\pi}_i$'s provide poor

approximations of the true $\hat{\pi}_i$'s. That is, GREG type estimators are model-consistent even if the sampling design (i.e., the distribution $F(\mathbf{I}|\mathbf{Z})$) is misspecified. Borrowing from the missing data literature, we say that GREG type estimators are doubly robust or doubly protected in the sense that they are design-consistent provided $F(\mathbf{I}|\mathbf{Z})$ is correctly specified, even if $F(\mathbf{y}|\mathbf{U})$ doesn't hold and it is model-consistent if $F(\mathbf{y}|\mathbf{U})$) holds irrespective of whether or not $F(\mathbf{I}|\mathbf{Z})$ is correctly specified. e.g., Kott and Liao (2012) and Kim and Haziza (2013). Clearly, GREG type estimators that incorporate the true inclusion probabilities π_i are automatically doubly robust. For rejective sampling (Fuller, 2009b), the π_i's are unknown. Fuller (2009b) suggested the use of a GREG type estimator based on the inclusion probabilities corresponding to the basic procedure, which are different of the inclusion probabilities π_i. As a result, the sampling design is misspecified. As we show empirically in Section 5, we argue that, although the GREG type estimator advocated by Fuller (2009b) is design-consistent, it may suffer from substantial bias for finite sample sizes when the p_i's do not provide a good approximation of the π_i's unless model (7.4) holds. On the other hand, since the Cube method ensure that the π_i's, the resulting Horvitz-Thompson estimator is exactly design unbiased, although it may suffer from instability if (7.4) does not hold.

7.2 The Cube algorithm

The cube method (Deville and Tillé, 2004) enables to select balanced samples (or approximately balanced samples) such that the inclusion probabilities $\boldsymbol{\pi}$ are exactly respected. The cube method proceeds in two steps: a flight phase, at the end of which an exact balancing is maintained, and a landing phase in which the balancing equations (7.1) may be partly relaxed until the complete sample is obtained, but the inclusion probabilities remain exactly respected.

The flight phase (Deville and Tillé, 2004; Chauvet and Tillé, 2006; Tillé, 2006) proceeds through a random walk from the vector of inclusion probabilities $\boldsymbol{\pi}$ to a random vector $\boldsymbol{\pi}^*$ such that $\pi_i^* = 0$ if unit i is definitely rejected from the sample, $\pi_i^* = 1$ if unit i is selected, and $0 < \pi_i^* < 1$ if the decision for unit i remains pending at the end of the flight phase. Denote by P^* the set of units such that $0 < \pi_i^* < 1$, so that $I_i = \pi_i^*$ for $i \notin P^*$. From Proposition 1 in Deville and Tillé (2004), the size of P^* is at most the number q of balancing variables. The flight phase is performed in such a way that

$$E\left(\boldsymbol{\pi}^*\right) = \boldsymbol{\pi}, \tag{7.5}$$

$$\sum_{i\in P} \frac{\mathbf{z}_i}{\pi_i}\pi_i^* = \sum_{i\in P} \mathbf{z}_i. \tag{7.6}$$

Equation (7.5) ensures that the inclusion probabilities are exactly respected at the end of the flight phase. Equation (7.6) ensures that the pseudo-estimator $\hat{t}_{\mathbf{x}\pi}^*$ is exactly balanced. Anyway, the flight phase does not lead to an estimator per se, since the selection is still not carried through for the units in P^*.

The landing phase enables to end the sampling, either by successively relaxing the balancing equations or by means of an enumerative algorithm on P^* (Tillé, 2006, p. 163). In any case, the landing phase is performed to obtain a vector of sample selection indicators $\mathbf{I}$ such that

$$\begin{aligned} E\left(\mathbf{I}\,|\boldsymbol{\pi}^*\right) &= \boldsymbol{\pi}^*, & (7.7)\\ \hat{t}_{\mathbf{z}}^{\pi} \equiv \sum_{i\in P} \frac{\mathbf{z}_i}{\pi_i} I_i &\simeq \sum_{i\in U} \mathbf{z}_i. & (7.8) \end{aligned}$$

Equation (7.7) ensures that the inclusion probabilities are exactly respected at the end of the landing phase, since from (7.5) and (7.7)

$$E(\mathbf{I}) = EE(\mathbf{I}|\boldsymbol{\pi}^*) = E(\boldsymbol{\pi}^*) = \boldsymbol{\pi}.$$

Consequently, the Horvitz-Thompson estimator $\hat{t}_y^{\pi}$ is exactly design-unbiased for t_y.

The non-exact balancing (see equation 7.8) results in an additional variability for the Horvitz-Thompson estimator. More precisely, the variance may be written as

$$\begin{aligned} V(\hat{t}_y^{\pi}) &= V\left[E\left(\hat{t}_y^{\pi}\,|\boldsymbol{\pi}^*\right)\right] + E\left[V\left(\hat{t}_y^{\pi}\,|\boldsymbol{\pi}^*\right)\right] \\ &= V\left[\sum_{i\in P}\frac{y_i}{\pi_i}\pi_i^*\right] + E\left[V\left(\sum_{i\in P^*}\frac{y_i}{\pi_i}I_i\middle|\,\boldsymbol{\pi}^*\right)\right]. \qquad (7.9) \end{aligned}$$

The first term in the right-hand side of (7.9) is the variance due to the flight phase. From equation (7.6), we have for any q-vector $\mathbf{B}$

$$\sum_{i\in P}\frac{y_i}{\pi_i}\pi_i^* = \sum_{i\in P}\frac{y_i - \mathbf{B}^\top\mathbf{z}_i}{\pi_i}\pi_i^* + \mathbf{B}^\top\sum_{i\in P}\mathbf{z}_i$$

which leads to

$$V_F(\hat{t}_y^{\pi}) \equiv V\left[\sum_{i\in P}\frac{y_i}{\pi_i}\pi_i^*\right] = V\left[\sum_{i\in P}\frac{E_i}{\pi_i}\pi_i^*\right] \qquad (7.10)$$

where $E_i = y_i - \mathbf{B}^\top\mathbf{z}_i$. The second term in the right-hand side of (7.9) is the variance due to the landing phase, and may be written as

$$V_L(\hat{t}_y^{\pi}) \equiv E\left[V\left(\sum_{i\in P^*}\frac{y_i}{\pi_i}I_i\middle|\,\boldsymbol{\pi}^*\right)\right] = E\left[\sum_{i,j\in P^*}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij}^* - \pi_i^*\pi_j^*)\right] \qquad (7.11)$$

where $\pi_{ij}^* = E(I_iI_j|\boldsymbol{\pi}^*)$. If the balancing variables have a large explanatory power for the variable y, the residuals E_i and the variance due to the flight phase are small (see equation 7.10), making the variance due the landing phase appreciable in the overall variance (Breidt and Chauvet, 2011). From equation (7.11), we also note that the variance due to the landing phase may be appreciable if the number of balancing variables is large as compared to the sample size (in which case, the random subpopulation P^* may be large as well) and/or if the inclusion probabilities π_i are poorly or negatively related to the variable y, so that the y_i/π_i are highly variable (see Chauvet, 2011).

7.3 Rejective sampling

Fuller (2009b) studied the theoretical properties of rejective sampling procedures that consists of discarding any sample that does not meet a specified balancing tolerance. Here, we adopt the notation of Fuller (2009b). The samples are selected by a sampling design called the basic procedure based on the vector of design variable $\mathbf{z}$, available for all $i \in P$. A subset of the $\mathbf{z}$-variables are the $\mathbf{z}_2$-variables, which are those that satisfy $V_b(\bar{\mathbf{z}}_2^p) = \mathbf{0}$, where $\bar{\mathbf{z}}_2^p = N^{-1}\sum_{i\in s_b} p_i^{-1}\mathbf{z}_{2i}$, is the vector of basic expansion estimators corresponding to $\mathbf{z}_2$. The design variables which are not part of the $\mathbf{z}_2$-variables are called the $\mathbf{z}_1$-variables. Let $\mathbf{x}_i = \mathbf{z}_{1i} - \mathbf{C}^\top \mathbf{z}_{2i}$, where $\mathbf{C}$ is the matrix that minimizes $\text{tr}\left\{V_p(\bar{\mathbf{z}}_1^p - \mathbf{C}^\top \bar{\mathbf{z}}_2^p)\right\}$ and $\bar{\mathbf{z}}_1^p = N^{-1}\sum_{i\in s_b} p_i^{-1}\mathbf{z}_{1i}$. Finally, let $\mathbf{z} = (\mathbf{x}^\top, \mathbf{z}_2^\top)$.

The rejective procedure proceeds as follows:

(i) select a sample, s_b, according to the basic procedure, using $p_1, \ldots, p_N$ as the vector of inclusion probabilities.

(ii) Let $\gamma > 0$ be a constant specified by the survey statistician. If (7.2) is satisfied, then the sample is retained, where $\hat{\mathbf{t}}_{\mathbf{x}}^p = \sum_{i\in s_b} p_i^{-1}\mathbf{x}_i$; otherwise, replace the sample in the population and repeat step (i).

A small value of γ corresponds to a high rejection rate. Recall that that $\pi_i \neq p_i$, in general. In fact, the π_i's are complex functions of $\hat{\mathbf{t}}_{\mathbf{x}}^\pi$ and $\mathbf{t}_{\mathbf{x}}$. As a result, they are generally untractable. Although condition (7.2) ensures that the sample s is approximately $p\mathbf{x}$-balanced for small values of γ, there is no guarantee that it is $\pi\mathbf{x}$-balanced.

We start by examining the bias of the basic expansion estimator $\hat{t}_y^p = \sum_{i\in s} p_i^{-1} y_i$ based on the basic inclusion probabilities p_i. Although $\hat{t}_y^p$ is design-unbiased for t_y with respect to the basic procedure, it is generally biased with respect to the rejective sampling procedure. The bias is given by

$$B_p(\hat{t}_y^p) = E_p(\hat{t}_y^p) - t_y = \sum_{i\in P} \delta_i y_i, \tag{7.12}$$

where $\delta_i = (\pi_i - p_i)/p_i$ and the subscript p denotes that the expectation is taken with respect to the rejective sampling design. The coefficient δ_i can be viewed as a measure of relative distance between the basic inclusion probabilities, p_i, and the inclusion probabilities with respect to the rejective sampling procedure, π_i. The bias in (7.12) is large if, for some units, the y-value is large and/or the δ-value is large. For simplicity, consider the case of a scalar x and suppose that the basic procedure is simple random sampling without replacement so that $p_i = n/N$ for all $i \in U$. We expect a large value of δ_i if x_i is much larger than the population mean $\bar{X} = t_x/N$. In this case, π_i is expected to be significantly smaller than p_i because most samples containing unit i are likely to be rejected, which in turns, leads to a large value of δ. Therefore, a unit exhibiting large values of x and y may have a significant contribution to the bias of $\hat{t}_y^p$. Finally, if the sample size is large enough, we expect the δ-values to be small as the distance between p_i and π_i becomes small as n increases.

To overcome this problem of bias, a possible option consists of estimating the π_i's through Monte Carlo simulations to obtain $\hat{\pi}_i^{MC}$; see Fattorini (2006) and Thompson and Wu (2008). Then, use the Monte Carlo expansion estimator

$$\hat{t}_y^{\hat{\pi}} = \sum_{i \in s} \frac{y_i}{\hat{\pi}_i^{MC}} \tag{7.13}$$

as an estimator of t_y. For large populations though, simulating enough samples to give precise estimates of the inclusion probabilities may be problematic.

An alternative option was studied in Fuller (2009b) who showed that the regression estimator based on the basic inclusion probabilities p_i and the vector of auxiliary variables $\mathbf{z}$, is design-consistent. More specifically, Fuller (2009b) advocated the use of the GREG type estimator

$$\hat{t}_{\text{reg}}^p = \sum_{i \in P} \mathbf{z}_i^\top \hat{\mathbf{B}}, \tag{7.14}$$

$$\hat{\mathbf{B}} = \left(\sum_{i \in s} p_i^{-2} \phi_i \mathbf{z}_i \mathbf{z}_i^\top \right)^{-1} \sum_{i \in s} p_i^{-2} \phi_i \mathbf{z}_i y_i$$

and ϕ_i is such that

$$E_b(\bar{\mathbf{x}}^p - \bar{\mathbf{X}} | i \in s_b) = p_i^{-1} N^{-1} \phi_i \mathbf{x}_i$$

with $\bar{\mathbf{x}}^p = N^{-1} \sum_{i \in s_b} p_i^{-1} \mathbf{x}_i$. For an arbitrary basic procedure, we have

$$E_b(\bar{\mathbf{x}}^p - \bar{\mathbf{X}} | i \in s_b) = \sum_{j \in U} \left(\frac{p_{ij} - p_i p_j}{p_i p_j} \right) \mathbf{x}_j,$$

where p_{ij} denotes the second-order inclusion probability of unit i and j in s_b. Fuller (2009b) showed that the GREG type estimator (7.14) is design-consistent provided that there exists a vector of constants $\boldsymbol{\lambda}$ such that

$$p_i^{-2} \phi_i \mathbf{z}_i^\top \boldsymbol{\lambda} = p_i^{-1}.$$

Example 4. *Consider the case of a scalar z_{1i}. If the basic procedure is simple random sampling without replacement with basic inclusion probabilities $p_i = n/N$ for all i, we have $z_{2i} = 1$ and $\mathbf{z}_i = (x_i, 1)^\top$ with $x_i = z_{1i} - \bar{Z}_1$ and $\bar{Z}_1 = N^{-1} \sum_{i \in P} z_{1i}$. Also, since*

$$E_b(\bar{\mathbf{x}}^p - \bar{\mathbf{X}} | i \in s_b) = 1/(N-1) \left(\frac{N}{n} - 1 \right) x_i,$$

we have $\phi_i = n/(N-1) \left(\frac{N}{n} - 1 \right)$ which does not depend on i.

Example 5. *Let $\mathbf{z}_{1i} = (1, z_{1i})^\top$. If the basic procedure is Bernoulli sampling with basic inclusion probabilities p, we have $z_{2i} = \emptyset$ and $\mathbf{z}_i = (x_i, 1)^\top$ with $x_i = z_{1i}$. Also, since*

$$E_b(\bar{\mathbf{x}}^p - \bar{\mathbf{X}} | i \in s_b) = N^{-1} \frac{(1-p)}{p} x_i,$$

we have $\phi_i = 1 - p$ which does not depend on i.

It is interesting to note that the GREG type estimator given by (7.14) can be alternatively expressed as

$$\hat{t}^p_{\text{reg}} = \sum_{i \in s} \frac{y_i}{\hat{\pi}^F_i}, \tag{7.15}$$

where

$$\hat{\pi}^F_i = \left\{ t_{\mathbf{z}}^\top \left(\sum_{i \in s} p_i^{-2} \phi_i \mathbf{z}_i \mathbf{z}_i^\top \right)^{-1} p_i^{-2} \phi_i \mathbf{z}_i \right\}^{-1}. \tag{7.16}$$

Thus, $\hat{\pi}^F_i$'s can be viewed as an (implicit) estimate of the true inclusion probability π_i. It turns out that for finite sample sizes, this approximation is generally not appropriate, especially if some units exhibit large values with respect to the balancing variables. This is illustrated empirically in Section 5. A better approximation of the true π_i's can be obtained through Edgeworth expansion as we illustrate in Section 4 for Poisson sampling as the basic procedure.

Using a first-order Taylor expansion, we obtain

$$\hat{t}^p_{\text{reg}} = \hat{t}^p_y + (\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{t}}^p_{\mathbf{z}})^\top \mathbf{B} + O_p(n^{-1}), \tag{7.17}$$

where

$$\mathbf{B} = \left(\sum_{i \in P} \pi_i p_i^{-2} \phi_i \mathbf{z}_i \mathbf{z}_i^\top \right)^{-1} \sum_{i \in P} \pi_i p_i^{-2} \phi_i \mathbf{z}_i y_i.$$

Ignoring the higher order terms in (7.17), the design-bias of $\hat{t}^p_{\text{reg}}$ can be approximated by

$$B_p(\hat{t}^p_{\text{reg}}) = \sum_{i \in P} \delta_i E_i, \tag{7.18}$$

where $E_i = y_i - \mathbf{z}_i^\top \mathbf{B}$. From (7.18), the bias of $\hat{t}^p_{\text{reg}}$ is small if the E_i's are small, which in turns indicates that model (7.4) holds. That is, there exists a strong linear relationship between y and $\mathbf{z}$. In multipurpose surveys, it is unrealistic to presume that (7.4) holds for every variable of interest. For example, some variables may be categorical rather than continuous, in which case model (7.4) may not provide a good description of the relationship between y and $\mathbf{z}$. In this case, the GREG type estimator (7.14) may be biased unless the sample size n is sufficiently large so that the δ-values are small.

In summary, under Fuller's estimation procedure, the sampling design is misspecified as it uses the basic inclusion probabilities p_i instead of the true inclusion probabilities π_i. In this case, one must put complete reliance on the model describing the relationship between y and $\mathbf{z}$. If the underlying model is linear, Fuller's procedure will exhibit low biases. However, as we show empirically in Section 5, large biases may occur in the case of non-linear relationships between y and $\mathbf{z}$.

7.4 Approximation of the inclusion probabilities through Edgeworth expansion

In this section, we obtain an approximation of the true inclusion probabilities π_i for Poisson sampling as the basic procedure. For simplicity, we consider the case of a scalar x_i. Condition (7.2) may be rewritten as $-\gamma \leq X \leq \gamma$ with $X = \frac{1}{\sqrt{d}} \sum_{i \in P} \tilde{x}_i (I_i - p_i)$, $d = \sum_{i \in P} \tilde{x}_i^2 \ p_i(1-p_i)$ and $\tilde{x}_i = x_i / p_i$. The final inclusion probability for rejective sampling is

$$\pi_i \ = \ p_i \frac{\mathbb{P}(-\gamma \leq X \leq \gamma | I_i = 1)}{\mathbb{P}(-\gamma \leq X \leq \gamma)}. \tag{7.19}$$

We first compute an expansion of $\mathbb{P}(-\gamma \leq X \leq \gamma) = F_X(\gamma) - F_X(-\gamma)$. From the formal Edgeworth expansion (see Thompson, 1997, equation (3.41) or Hàjek, 1964 et 1981), we get

$$\begin{aligned} F_X(\gamma) \ &= \ \psi(\gamma) - \phi(\gamma) \left\{ \frac{\kappa_3}{6} (\gamma^2 - 1) \right\} \\ &\quad - \phi(\gamma) \left\{ \frac{\kappa_4}{24} (\gamma^3 - 3\gamma) + \frac{\kappa_3^2}{72} (\gamma^5 - 10\gamma^3 + 15\gamma) \right\} + o(d^{-1}), \\ F_X(-\gamma) \ &= \ 1 - \psi(\gamma) - \phi(\gamma) \left\{ \frac{\kappa_3}{6} (\gamma^2 - 1) \right\} \\ &\quad + \phi(\gamma) \left\{ \frac{\kappa_4}{12} (\gamma^3 - 3\gamma) + \frac{\kappa_3^2}{36} (\gamma^5 - 10\gamma^3 + 15\gamma) \right\} + o(d^{-1}), \end{aligned} \tag{7.20}$$

where $\psi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and the probability density function of a standard normal distribution, where

$$\begin{aligned} \kappa_3 \ &\equiv \ \mu_3(X) = d^{-3/2} \sum_{i \in P} \tilde{x}_i^3 \ p_i(1-p_i)(1-2p_i), \\ \kappa_4 \ &\equiv \ \mu_4(X) - 3(\mu_2(X))^2 = d^{-2} \sum_{i \in P} \tilde{x}_i^4 \ p_i(1-p_i)(1-6p_i(1-p_i)), \end{aligned}$$

and $\mu_m(X)$ denotes the centered moment of order m of the random variable X. Equations (7.20) lead to

$$\begin{aligned} \mathbb{P}(-\gamma \leq X \leq \gamma) \ &= \ \{2\psi(\gamma) - 1\} \\ &\quad - \phi(\gamma) \left\{ \frac{\kappa_4}{12} (\gamma^3 - 3\gamma) + \frac{\kappa_3^2}{36} (\gamma^5 - 10\gamma^3 + 15\gamma) \right\} + o(d^{-1}). \end{aligned} \tag{7.21}$$

We now compute an expansion of $\mathbb{P}(-\gamma \leq X \leq \gamma | I_i = 1)$. We have

$$\mathbb{P}(X \leq \gamma | I_i = 1) = \mathbb{P}(X_i \leq \gamma_i),$$

where

$$\begin{aligned} X_i \ &= \ \frac{1}{\sqrt{d - \tilde{x}_i^2 \ p_i(1-p_i)}} \sum_{j \neq i \in U} \tilde{x}_j (I_j - p_j), \\ \gamma_i \ &= \ \frac{1}{\sqrt{1 - d^{-1} \tilde{x}_i^2 \ p_i(1-p_i)}} \left(\gamma - d^{-1/2} \tilde{x}_i \ (1 - p_i) \right). \end{aligned}$$

Applying once again the formal Edgeworth expansion to X_i leads to

$$\begin{aligned} F_{X_i}(\gamma_i) &= \psi(\gamma_i) - \phi(\gamma_i)\left\{\frac{\kappa_{3i}}{6}(\gamma_i^2-1)\right\} \\ &\quad - \phi(\gamma_i)\left\{\frac{\kappa_{4i}}{24}(\gamma_i^3-3\gamma_i) + \frac{\kappa_{3i}^2}{72}(\gamma_i^5-10\gamma_i^3+15\gamma_i)\right\} + o(d^{-1}), \end{aligned} \tag{7.22}$$

where

$$\begin{aligned} \kappa_{3i} &\equiv \mu_3(X_i) = \left(d - \tilde{x}_i^2\ p_i(1-p_i)\right)^{-3/2} \sum_{j\neq i\in U} \tilde{x}_j^3\ p_j(1-p_j)(1-2p_j), \\ \kappa_{4i} &\equiv \mu_4(X_i) - 3(\mu_2(X_i))^2 = \left(d - \tilde{x}_i^2\ p_i(1-p_i)\right)^{-2} \sum_{j\neq i\in U} \tilde{x}_j^4\ p_j(1-p_j)(1-6p_j(1-p_j)). \end{aligned}$$

By neglecting terms of smaller order than d^{-1}, we obtain after some algebra

$$\begin{aligned} F_{X_i}(\gamma_i) &= \psi(\gamma) - \phi(\gamma)\left\{\frac{1}{\sqrt{d}}\tilde{x}_i(1-p_i) + \frac{1}{6}\kappa_3(\gamma^2-1)\right\} \\ &\quad + \left\{\frac{1}{2d}\tilde{x}_i^2(1-p_i)\left(\gamma p_i\phi(\gamma) + (1-p_i)\phi'(\gamma)\right)\right. \\ &\quad\quad + \frac{\kappa_3\gamma\phi(\gamma)}{3\sqrt{d}}\tilde{x}_i(1-p_i) - \frac{\kappa_4\phi(\gamma)}{24}(\gamma^3-3\gamma) \\ &\quad\quad \left. -\phi(\gamma)\frac{\kappa_3^2}{72}(\gamma^5-10\gamma^3+15\gamma) + \frac{1}{6}\frac{\kappa_3(\gamma^2-1)}{\sqrt{d}}\tilde{x}_i(1-p_i)\phi'(\gamma)\right\} \\ &\quad + o(d^{-1}). \end{aligned} \tag{7.23}$$

Also, $\mathbb{P}(X \leq -\gamma | I_i = 1) = F_{X_i}(-\tilde{\gamma}_i)$ where

$$\tilde{\gamma}_i = \frac{1}{\sqrt{1 - d^{-1}\tilde{x}_i^2\ p_i(1-p_i)}}\left(\gamma + d^{-1/2}\tilde{x}_i\ (1-p_i)\right).$$

Similar arguments lead to

$$\begin{aligned} F_{X_i}(-\tilde{\gamma}_i) &= \{1-\psi(\gamma)\} - \phi(\gamma)\left\{\frac{1}{\sqrt{d}}\tilde{x}_i(1-p_i) + \frac{1}{6}\kappa_3(\gamma^2-1)\right\} \\ &\quad + \left\{\frac{1}{2d}\tilde{x}_i^2(1-p_i)\left(-\gamma p_i\phi(\gamma) - (1-p_i)\phi'(\gamma)\right)\right. \\ &\quad\quad - \frac{\kappa_3\gamma\phi(\gamma)}{3\sqrt{d}}\tilde{x}_i(1-p_i) + \frac{\kappa_4\phi(\gamma)}{24}(\gamma^3-3\gamma) \\ &\quad\quad \left. +\phi(\gamma)\frac{\kappa_3^2}{72}(\gamma^5-10\gamma^3+15\gamma) - \frac{1}{6}\frac{\kappa_3(\gamma^2-1)}{\sqrt{d}}\tilde{x}_i(1-p_i)\phi'(\gamma)\right\} \\ &\quad + o(d^{-1}). \end{aligned} \tag{7.24}$$

We obtain

$$\begin{aligned}\mathbb{P}(-\gamma \leq X \leq \gamma | I_i = 1) &= \{2\psi(\gamma) - 1\} + \frac{1}{d}\tilde{x}_i^2(1-p_i)\left(\gamma p_i \phi(\gamma) + (1-p_i)\phi'(\gamma)\right) \quad (7.25)\\ &+ 2\left\{\frac{\kappa_3 \gamma \phi(\gamma)}{3\sqrt{d}}\tilde{x}_i(1-p_i) - \frac{\kappa_4 \phi(\gamma)}{24}(\gamma^3 - 3\gamma)\right.\\ &\left. -\phi(\gamma)\frac{\kappa_3^2}{72}(\gamma^5 - 10\gamma^3 + 15\gamma) + \frac{1}{6}\frac{\kappa_3(\gamma^2-1)}{\sqrt{d}}\tilde{x}_i(1-p_i)\phi'(\gamma)\right\}\\ &+ o(d^{-1}).\end{aligned}$$

By plugging (7.21) and (7.25) into (7.19), and making use of the identity $\phi'(\gamma) = -\gamma\phi(\gamma)$, this leads to

$$\begin{aligned}\pi_i &= p_i\left\{1 - \frac{1}{d}\frac{\gamma\phi(\gamma)}{2\psi(\gamma)-1}\tilde{x}_i^2(1-p_i)(1-2p_i) \right. \quad (7.26)\\ &\left. + \frac{\kappa_3}{\sqrt{d}}\frac{\gamma\phi(\gamma)(3-\gamma^2)}{3(2\psi(\gamma)-1)}\tilde{x}_i(1-p_i)\right\} + o(d^{-1}).\end{aligned}$$

Assuming that $\gamma = O(d^{-0.5})$, this approximation simplifies to

$$\pi_i = p_i\left\{1 - \frac{1}{2d}\tilde{x}_i^2(1-p_i)(1-2p_i) + \frac{\kappa_3}{2\sqrt{d}}\tilde{x}_i(1-p_i)\right\} + o(d^{-1}). \quad (7.27)$$

Comparing (7.16) and (7.27) helps understanding why the GREG type estimator (7.14) advocated by Fuller (2009b) may perform poorly as both expression are significantly different.

7.5 Simulation Study

We conducted an extensive simulation study in order to compare the performance of several estimators in terms of relative bias and mean square error. We generated several populations of size $N = 500$, each consisting of a variable x and a variable of interest y. In each population, the x-values were first generated according to three distributions:
(1) normal distribution: $x_i \sim \mathcal{N}(2,1)$ for $i \in (1,\ldots,N)$;
(2) mixture distribution: $x_i \sim \mathcal{N}(2,1)$ for $i \in (6,\ldots,N)$ and the first five values $x_1 - x_5$ were manually set to $x_1 = 8$, $x_2 = 8.2$, $x_3 = 7.9$, $x_4 = 8.3$ and $x_5 = 8.5$;
(3) lognormal distribution: $\ln(x_i) \sim \mathcal{N}(0, 0.9)$ for $i \in (1,\ldots,N)$.
For the mixture distribution, note that the $x_1 - x_5$-values are larger than the remaining of the x-values. Similarly, because of the asymmetric nature of the lognormal distribution, some units. exhibited large x-values with respect to the others.

Given the x-values, the y-values were generated according to 4 superpopulation models:
(1) linear regression model: $y_i = 1 + 2(x_i - \bar{X}) + \varepsilon_i^{(1)}$, where $\varepsilon_i^{(1)} \sim \mathcal{N}(0,1)$;
(2) quadratic regression model: $y_i = 1 + 2(x_i - \bar{X})^2 + \varepsilon_i^{(2)}$, where $\varepsilon_i^{(2)} \sim \mathcal{N}(0,1)$;
(3) exponential model: $y_i = \exp(1 + x_i - \bar{X})$;

(4) bump model: $y_i = 1 + 2(x_i - \bar{X})^2 - 10\exp\left\{-20(x_i - \bar{X})^2\right\} + \varepsilon_i^{(4)}$, where $\varepsilon_i^{(4)} \sim \mathcal{N}(0,1)$.

This led to the creation of 12 populations. Form each population, we selected samples according to the rejective procedure of Fuller (2009b). We used two basic sampling procedures: (i) simple random sampling without replacement and (ii) Bernoulli sampling. We used three samples sizes: $n \in (25,\ 50,\ 75)$, where n is the effective sample size in the case of simple random sampling without replacement and is the expected sample size in the case of Bernoulli sampling. We used a rejective rule, which is slightly different from the one described in Section 3. Samples were selected using the basic sampling procedure until

$$\left|\frac{\hat{t}_{x,\pi} - t_x}{t_x}\right| < \tau,$$

where τ is balancing tolerance. In our view the above rule is more natural from a practical point of view. Although the balancing tolerances γ (see Section 3) and τ are different, it is straightforward to determine the value τ for a given value of γ and vice versa. We used three values of τ: $0.1\%, 1\%$ and 5%.

In each sample, we computed 4 linear estimators: (a) the basic expansion estimator BEE, $\hat{t}_y^p = \sum_{i\in s} p_i^{-1} y_i$; (b) the Monte Carlo expansion estimator (MC), given by (7.13) (c) the Fuller's estimator (Fuller) given by (7.14) and (d) the Edgeworth expansion estimator (Edge.), $\sum_{i\in s} y_i/\hat{\pi}_i^{Edge}$, where $\hat{\pi}_i^{Edge}$ is given by (7.27), excluding the $o(d^{-1})$ term.

For the Monte Carlo expansion estimator (b) approximations of the π_i's resulting from the rejective procedure were separately obtained through
$K_1 = 500\ 000$ Monte Carlo runs.

To assess the bias and efficiency of the four estimators, we used $K_2 = 10\ 000$ Monte Carlo runs. As the weights $1/\hat{\pi}^F$ in Fuller's estimator vary from one sample to the other, we also computed their Monte Carlo average over the the K_2 sets of weights.

For an estimator $\hat{t}$, we computed its Monte Carlo percent relative bias (in %) given by

$$RB_{MC}\left(\hat{t}\right) = \frac{1}{K_2}\sum_{j=1}^{K_2}\frac{(\hat{t}_{(j)} - t_y)}{t_y} \times 100,$$

where $\hat{t}_{(j)}$ denotes the estimator $\hat{t}$ at the j-th iteration.

As a measure of variability of $\hat{t}$, we computed its Monte Carlo coefficient of variation (in %) given by

$$CV_{MC}(\hat{t}) = 100 \times \frac{\left[\frac{1}{K_2}\sum_{j=1}^{K_2}\left(\hat{t}_{(j)} - t_y\right)^2 - \left\{\frac{1}{K_2}\sum_{j=1}^{K_2}(\hat{t}_{(j)} - t_y)\right\}^2\right]^{0.5}}{t_y}.$$

As a measure of efficiency of $\hat{t}$, using Fuller's estimator $\hat{t}^p_{reg}$ as the reference, we used

$$RRMSE_{MC}(\hat{t}) = 100 \times \frac{\left\{\frac{1}{K_2}\sum_{j=1}^{K_2}\left(\hat{t}_{(j)} - t_y\right)^2\right\}^{0.5}}{\left\{\frac{1}{K_2}\sum_{j=1}^{K_2}\left(\hat{t}^p_{reg(j)} - t_y\right)^2\right\}^{0.5}},$$

Tables (7.1)-(7.12) show the values of the Monte Carlo results in terms of relative bias, coefficient of variation and relative root mean square error of the 4 estimators.

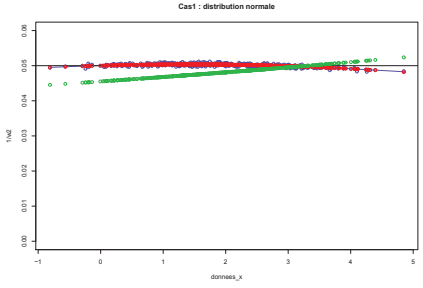
Tables (7.1)-(7.12) show the values of the Monte Carlo results in terms of relative bias, coefficient of variation and relative root mean square error of the 4 estimators.

Figures(7.5)-(7.3) plot the different inclusion probabilities corresponding to each of the 4 estimators. The flat line in black corresponds to the basic inclusion probabilities, which are all equal as the basic procedures are simple random sampling without replacement and Bernoulli sampling. The blue dots represent the Monte Carlo approximation of the π_i's and the blue curve is a smoothed adjustment curve. The red dots represent the approximation of the π_i's obtained through Edgeworth expansion; see expression (7.27) derived for Poisson sampling. Finally, the green dots represent the Monte Carlo average of $\hat{\pi}_i^F$ in Fuller's estimator.

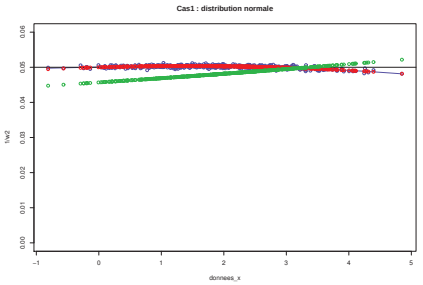
We first discuss the figures (7.5)-(7.3). From the Monte Carlo approximation of the π_i's, we note that the units exhibiting large x-values have an inclusion probability π_i significantly different from the basic inclusion probability p_i. This is especially apparent when the x-values were generated from a mixture distribution and from a lognormal distribution. For Bernoulli sampling, we note that the Monte Carlo approximation of the π_i's led to almost identical results obtained through Edgeworth expansion. On the other hand, we note that the inclusion probabilities $\hat{\pi}_i^F$ involved in Fuller's estimator are often poor approximation of the true π_i's, especially for large x-values. This is particularly apparent when the x-values were generated from a mixture distribution and from a lognormal distribution. As the (expected) sample size increased, we note that all the approximations of the $\pi - i$'s get closer to the basic inclusion probabilities p_i, as expected. Also, we note that the value of the balancing tolerance τ did not seem to make much difference. Finally, Figure (7.3) shows the approximation of the π_i obtained through Edgeworth expansion for simple random sampling without replacement is not very good. This result is not surprising as (7.27) was derived for Bernoulli sampling and not for simple random sampling without replacement.

We now discuss the results shown in Tables (7.1)-(7.12). As expected, the basic expansion estimator was generally biased. The bias was larger when the x-variables was generated according to a mixture distribution or a lognormal. The Monte Carlo expansion estimator showed small biases in all the scenarios. For Bernoulli sampling, the Edgeworth expansion estimator and the Monte Carlo expansion estimator showed almost identical results. Fuller's estimator showed negligible biases when the y-values were generated from a linear model, $y(1)$, as expected. However, it showed

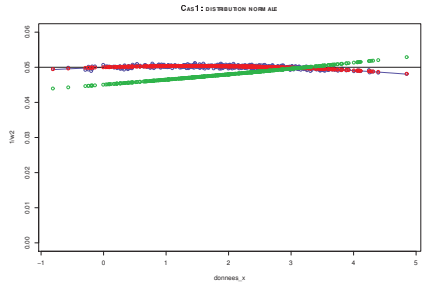
large biases when the y-variable was generated according to the non-linear model, $y(2)$, $y(3)$ and $y(4)$. In fact, the bias of Fuller's estimator was often superior to that of the basic expansion estimator, illustrating the risks of using Fuller's estimator in a multipurpose survey. In terms of efficiency, Fuller's estimator exhibited the smallest coefficient of variation in the majority of the scenarios. However, when biased, its mean square error was often superior to that of the Monte Carlo expansion estimator or the Edgeworth expansion estimator.



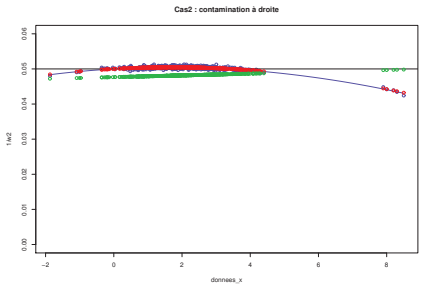
(a) $x(1)$ - $\tau = 0.1\%$



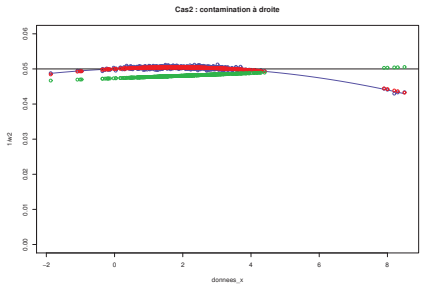
(b) $x(1)$ - $\tau = 1\%$



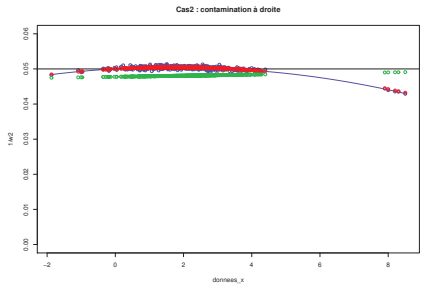
(c) $x(1)$ - $\tau = 5\%$



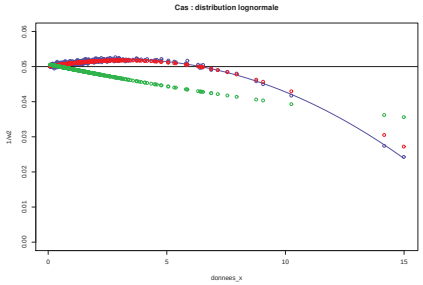
(d) $x(2)$ - $\tau = 0.1\%$



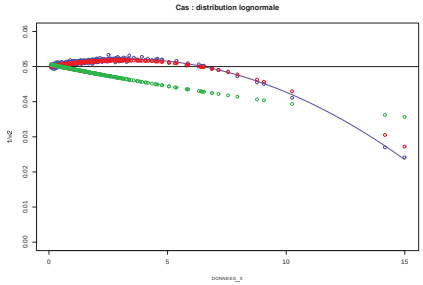
(e) $x(2)$ - $\tau = 1\%$



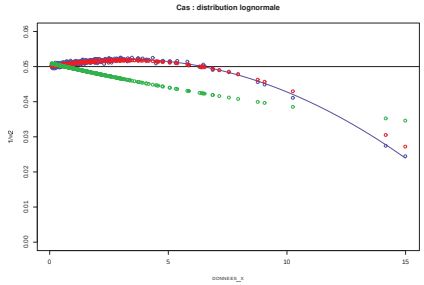
(f) $x(2)$ - $\tau = 5\%$



(g) $x(3)$ - $\tau = 0.1\%$

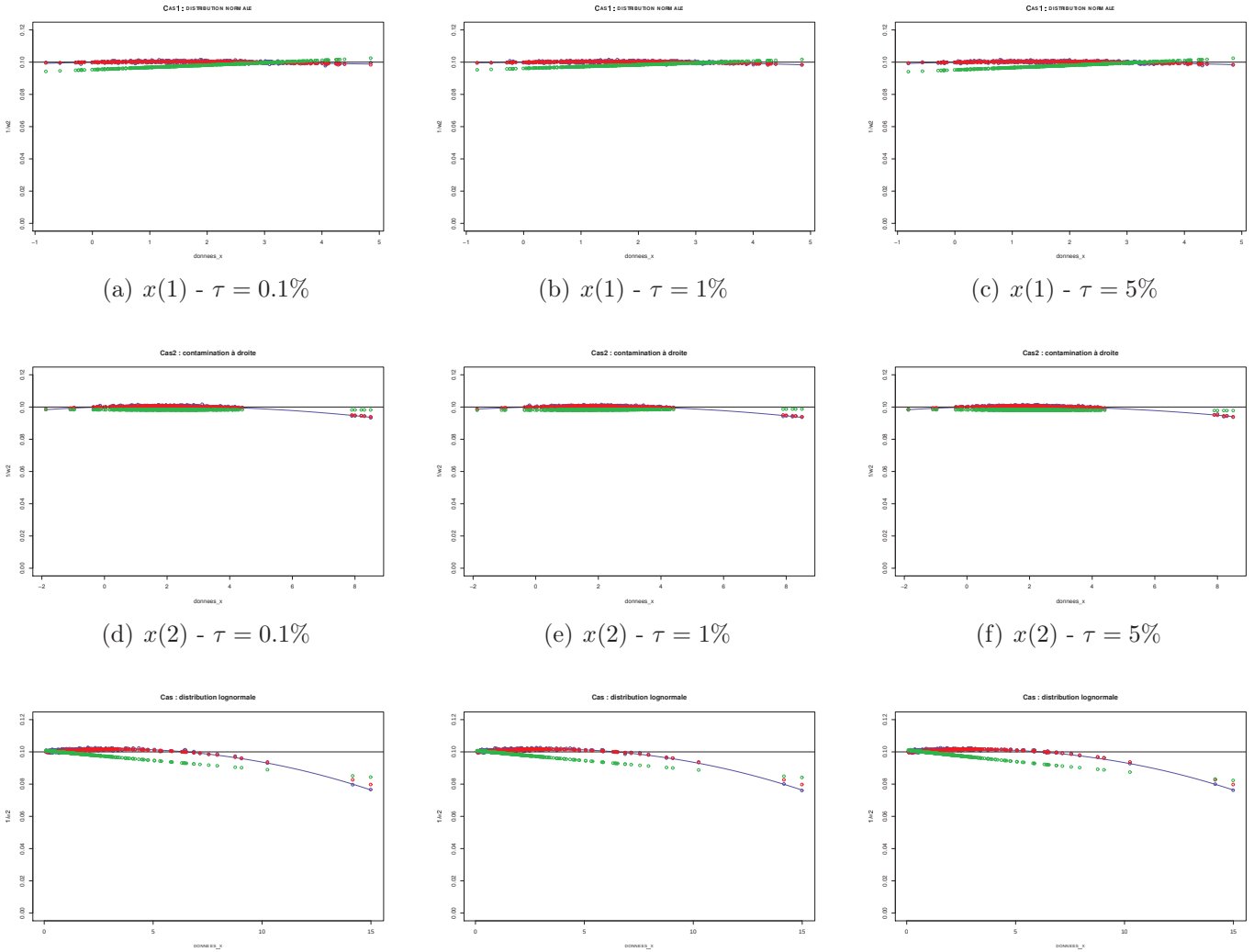


(h) $x(3)$ - $\tau = 1\%$



(i) $x(3)$ - $\tau = 5\%$

FIGURE 7.1 – Inverse of the weights of the 4 estimators for a rejective Bernoulli sampling of size $n = 25$; BEE (in black), MC (in blue), Fuller (in green) and Edge. (in red).

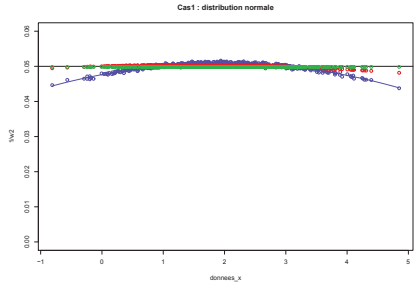


(a) $x(1)$ - $\tau = 0.1\%$ (b) $x(1)$ - $\tau = 1\%$ (c) $x(1)$ - $\tau = 5\%$

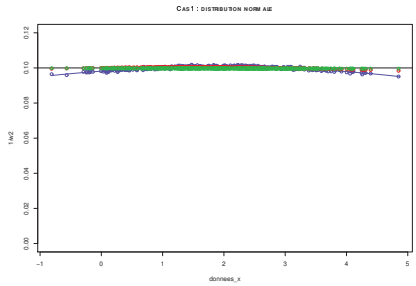
(d) $x(2)$ - $\tau = 0.1\%$ (e) $x(2)$ - $\tau = 1\%$ (f) $x(2)$ - $\tau = 5\%$

(g) $x(3)$ - $\tau = 0.1\%$ (h) $x(3)$ - $\tau = 1\%$ (i) $x(3)$ - $\tau = 5\%$

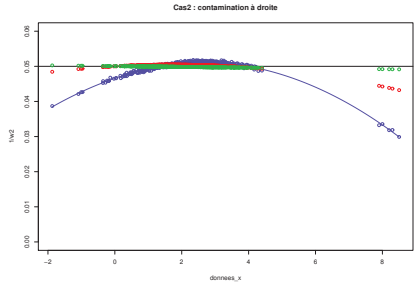
FIGURE 7.2 – Inverse of the weights of the 4 estimators for a rejective Bernoulli sampling of size $n = 50$; BEE (in black), MC (in blue), Fuller (in green) and Edge. (in red).



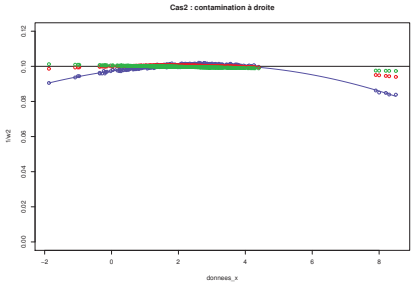
(a) x(1) - n=25



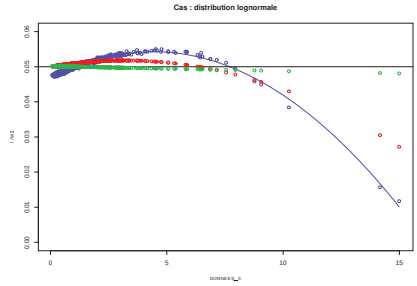
(b) x(1) - n=50



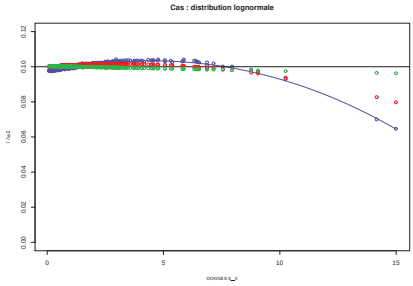
(c) x(2) - n=25



(d) x(2) - n=50



(e) x(3) - n=25



(f) x(3) - n=50

FIGURE 7.3 – Inverse of the weights of the 4 estimators for a rejective simple random sampling with $\tau = 5\%$; BEE (in black), MC (in blue), Fuller (in green) and Edge. (in red).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-1.06	-0.44	0.08	-0.49	-0.08	0.29	-0.04	0.28	-0.08	0.13	-0.08	0.09
	CV	30.53	30.59	19.29	30.62	26.83	26.86	15.16	26.87	17.80	17.82	10.57	17.82
	RRMSE	158.41	158.65	100.00	158.77	177.00	177.23	100.00	177.29	168.38	168.57	100.00	168.53
$y(2)$-Quadratic	RB	0.22	0.37	-3.39	0.38	-0.21	-0.14	-1.80	-0.14	0.02	0.05	-0.96	0.06
	CV	19.31	19.41	17.78	19.42	13.61	13.65	12.41	13.65	10.70	10.72	9.65	10.72
	RRMSE	106.71	107.28	100.00	107.33	108.59	108.83	100.00	108.86	110.33	110.45	100.00	110.50
$y(3)$-Exponential	RB	-0.48	0.01	-2.49	-0.01	-0.22	-0.03	-1.28	-0.01	0.05	0.17	-0.64	0.17
	CV	14.85	15.32	13.05	15.33	10.31	10.41	8.89	10.45	8.22	8.28	7.12	8.28
	RRMSE	111.78	115.27	100.00	115.36	114.74	115.89	100.00	116.35	114.86	115.81	100.00	115.84
$y(4)$-Bump	RB	-0.40	0.57	-10.47	0.52	-0.73	-0.40	-4.79	-0.37	-0.15	0.01	-2.69	0.07
	CV	59.77	59.80	62.79	59.84	37.55	37.56	38.13	37.57	29.71	29.72	29.93	29.72
	RRMSE	93.90	93.94	100.00	94.01	97.74	97.75	100.00	97.76	98.88	98.92	100.00	98.91

TABLE 7.1 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 0.1\%$ and balancing variable x(1-Normal distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-1.56	-0.30	0.19	-0.34	-0.78	-0.11	-0.18	-0.17	-0.31	0.10	0.18	0.05
	CV	33.65	34.89	19.37	34.81	24.56	24.95	13.18	24.93	21.09	21.24	11.07	21.24
	RRMSE	173.87	180.09	100.00	179.70	186.40	189.20	100.00	189.10	190.46	191.82	100.00	191.79
$y(2)$-Quadratic	RB	-2.43	-0.13	-12.17	-0.22	-0.90	0.18	-6.58	0.13	-0.87	-0.27	-4.79	-0.28
	CV	34.92	39.28	32.57	39.13	25.19	26.55	24.89	26.47	20.24	20.84	20.40	20.81
	RRMSE	100.69	112.99	100.00	112.55	97.91	103.10	100.00	102.81	96.67	99.47	100.00	99.32
$y(3)$-Exponential	RB	-10.67	-1.26	-35.26	-1.60	-3.11	1.33	-19.41	1.09	-2.91	-0.38	-14.34	-0.52
	CV	130.96	151.11	103.13	150.10	95.16	101.25	82.88	100.91	75.62	78.41	67.62	78.21
	RRMSE	120.55	138.65	100.00	137.72	111.85	118.96	100.00	118.55	109.47	113.43	100.00	113.14
$y(4)$-Bump	RB	-4.43	-0.46	-20.65	-0.64	-1.43	0.37	-10.58	0.31	-1.69	-0.69	-7.78	-0.71
	CV	60.63	66.71	58.21	66.51	42.89	44.77	42.51	44.66	33.37	34.20	33.46	34.15
	RRMSE	98.42	108.01	100.00	107.69	97.97	102.21	100.00	101.96	97.26	99.57	100.00	99.45

TABLE 7.2 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 0.1\%$ and balancing variable x(2-Mixtured distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-3.51	0.19	0.09	-0.68	-1.83	-0.22	0.12	-0.43	-1.06	-0.14	-0.29	-0.22
	CV	37.09	48.70	20.66	45.45	25.81	27.90	13.73	27.50	19.43	20.25	10.91	20.12
	RRMSE	180.30	235.69	100.00	219.96	188.44	203.20	100.00	200.25	178.33	185.59	100.00	184.41
$y(2)$-Quadratic	RB	-10.95	-0.68	-17.09	-2.80	-4.89	-0.35	-8.85	-1.02	-2.57	0.16	-5.50	-0.20
	CV	45.37	84.82	38.01	76.05	34.70	44.39	30.07	42.71	28.30	32.42	25.16	31.83
	RRMSE	112.01	203.56	100.00	182.63	111.81	141.63	100.00	136.32	110.35	125.91	100.00	123.58
$y(3)$-Exponential	RB	-51.24	-4.54	-66.01	-14.53	-24.07	-2.42	-40.72	-6.12	-12.65	0.23	-26.63	-1.67
	CV	224.35	453.27	166.70	404.74	192.41	249.67	160.36	239.99	162.01	187.01	144.96	183.24
	RRMSE	128.35	252.82	100.00	225.89	117.20	150.91	100.00	145.10	110.26	126.89	100.00	124.33
$y(4)$-Bump	RB	-13.58	-0.93	-21.01	-3.45	-6.02	-0.34	-10.81	-1.15	-3.27	0.23	-6.90	-0.21
	CV	56.96	102.61	47.83	92.37	43.83	55.17	37.71	53.20	36.70	41.63	32.11	40.91
	RRMSE	112.07	196.40	100.00	176.92	112.78	140.66	100.00	135.66	112.20	126.77	100.00	124.58

TABLE 7.3 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 0.1\%$ and balancing variable x(3-Lognormal distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-1.30	-0.61	0.05	-0.63	-0.60	-0.32	0.29	-0.29	-0.11	0.07	-0.00	0.05
	CV	34.84	34.94	21.94	34.93	25.76	25.80	15.21	25.79	17.09	17.10	10.48	17.11
	RRMSE	158.89	159.28	100.00	159.23	169.37	169.58	100.00	169.54	163.07	163.21	100.00	163.22
$y(2)$-Quadratic	RB	0.25	0.45	-3.60	0.44	0.04	0.15	-1.78	0.12	-0.04	-0.00	-1.11	0.00
	CV	20.51	20.60	19.08	20.62	13.87	13.91	12.62	13.90	10.97	10.98	9.96	10.99
	RRMSE	105.66	106.16	100.00	106.24	108.85	109.14	100.00	109.12	109.44	109.56	100.00	109.59
$y(3)$-Exponential	RB	-0.43	0.03	-2.36	0.04	-0.27	-0.05	-1.19	-0.06	-0.17	-0.03	-0.80	-0.04
	CV	14.79	15.27	13.14	15.28	10.24	10.39	8.88	10.38	8.29	8.37	7.10	8.36
	RRMSE	110.87	114.39	100.00	114.44	114.33	115.96	100.00	115.92	116.07	117.13	100.00	117.01
$y(4)$-Bump	RB	0.20	1.02	-9.44	1.06	-0.25	0.20	-4.59	0.13	-0.11	0.09	-2.74	0.10
	CV	56.73	56.78	59.19	56.80	38.17	38.18	38.76	38.19	29.73	29.73	29.96	29.74
	RRMSE	94.66	94.76	100.00	94.78	97.80	97.84	100.00	97.85	98.81	98.82	100.00	98.84

TABLE 7.4 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 1\%$ and balancing variable x(1-Normal distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-1.31	0.06	0.13	0.01	-0.75	-0.12	-0.20	-0.15	-0.42	-0.08	0.04	-0.05
	CV	35.54	36.76	19.60	36.70	24.69	25.02	13.48	25.00	20.16	20.32	10.74	20.33
	RRMSE	181.42	187.56	100.00	187.24	183.30	185.67	100.00	185.53	187.83	189.31	100.00	189.34
$y(2)$-Quadratic	RB	0.25	0.45	-3.60	0.44	0.04	0.15	-1.78	0.12	-0.04	-0.00	-1.11	0.00
	CV	36.44	41.08	33.38	40.88	25.53	26.90	25.30	26.81	20.00	20.56	20.28	20.57
	RRMSE	102.84	115.71	100.00	115.16	97.24	102.26	100.00	101.93	95.83	98.39	100.00	98.44
$y(3)$-Exponential	RB	-8.99	0.79	-33.93	0.35	-6.16	-1.89	-21.33	-2.16	-3.61	-1.30	-14.90	-1.25
	CV	134.34	154.86	104.35	154.04	92.61	98.53	81.48	98.20	75.13	77.64	67.62	77.71
	RRMSE	122.71	141.14	100.00	140.38	110.19	117.00	100.00	116.61	108.63	112.15	100.00	112.24
$y(4)$-Bump	RB	-3.46	0.58	-19.44	0.45	-2.68	-0.94	-11.31	-1.03	-1.91	-0.97	-8.16	-0.92
	CV	62.08	68.51	58.40	68.23	41.95	43.74	41.86	43.63	34.06	34.83	34.32	34.85
	RRMSE	101.02	111.32	100.00	110.86	96.96	100.91	100.00	100.65	96.73	98.79	100.00	98.83

TABLE 7.5 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 1\%$ and balancing variable x(2-Mixtured distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-3.44	0.37	-0.04	-0.60	-1.65	-0.31	0.30	-0.52	-0.94	-0.02	0.14	-0.08
	CV	34.90	47.77	20.20	43.93	24.66	26.58	13.47	26.18	20.49	21.25	11.00	21.14
	RRMSE	173.59	236.48	100.00	217.48	183.44	197.29	100.00	194.38	186.48	193.19	100.00	192.18
$y(2)$-Quadratic	RB	-10.58	0.53	-17.06	-2.04	-5.71	-1.40	-9.66	-2.07	-2.61	0.11	-5.40	-0.24
	CV	46.16	87.83	37.68	77.89	34.09	43.49	29.38	41.80	28.21	32.39	25.26	31.76
	RRMSE	114.51	212.38	100.00	188.39	111.75	140.70	100.00	135.31	109.67	125.41	100.00	122.95
$y(3)$-Exponential	RB	-49.01	0.74	-65.62	-10.58	-29.40	-8.95	-45.34	-12.70	-12.08	1.13	-25.17	-1.00
	CV	229.39	466.56	165.35	413.89	186.76	243.78	154.94	232.99	163.30	189.19	147.94	184.73
	RRMSE	131.86	262.27	100.00	232.73	117.11	151.11	100.00	144.54	109.12	126.08	100.00	123.11
$y(4)$-Bump	RB	-13.50	0.56	-21.45	-2.59	-7.34	-1.87	-12.17	-2.70	-3.13	0.22	-6.48	-0.20
	CV	59.61	109.19	48.60	97.26	43.59	54.67	37.31	52.67	35.10	39.95	31.12	39.21
	RRMSE	115.07	205.58	100.00	183.16	112.63	139.39	100.00	134.38	110.86	125.66	100.00	123.34

TABLE 7.6 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 1\%$ and balancing variable x(3-Lognormal distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-0.53	0.05	0.14	0.03	-0.15	0.20	-0.09	0.13	-0.37	-0.13	-0.11	-0.19
	CV	30.03	30.12	18.44	30.11	21.06	21.09	12.28	21.09	18.98	19.00	10.89	19.00
	RRMSE	162.93	163.39	100.00	163.35	171.42	171.66	100.00	171.64	174.35	174.51	100.00	174.48
$y(2)$-Quadratic	RB	0.04	0.29	-3.68	0.24	-0.04	0.08	-1.61	0.04	-0.15	-0.08	-1.15	-0.10
	CV	20.14	20.30	18.94	20.28	14.20	14.24	12.79	14.24	11.05	11.06	9.90	11.07
	RRMSE	104.43	105.24	100.00	105.14	110.15	110.48	100.00	110.43	110.89	111.02	100.00	111.06
$y(3)$-Exponential	RB	-0.36	0.16	-2.53	0.11	-0.16	0.08	-1.18	0.05	-0.25	-0.09	-0.78	-0.13
	CV	15.18	15.68	13.20	15.66	10.84	10.98	9.02	10.98	8.82	8.89	7.04	8.89
	RRMSE	112.97	116.67	100.00	116.53	119.06	120.69	100.00	120.64	124.52	125.40	100.00	125.40
$y(4)$-Bump	RB	-0.30	0.63	-9.89	0.57	-0.45	-0.02	-4.75	-0.06	-0.33	-0.16	-2.88	-0.12
	CV	56.85	56.95	60.00	56.91	39.70	39.73	40.43	39.73	29.78	29.79	29.87	29.79
	RRMSE	93.48	93.66	100.00	93.59	97.53	97.60	100.00	97.58	99.25	99.30	100.00	99.28

TABLE 7.7 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 5\%$ and balancing variable x(1-Normal distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-2.27	-0.49	-0.03	-0.69	-0.68	-0.01	0.03	-0.09	-0.47	-0.04	-0.03	-0.09
	CV	40.21	41.63	21.60	41.57	24.92	25.27	13.22	25.27	20.92	21.11	10.23	21.10
	RRMSE	186.44	192.77	100.00	192.48	188.53	191.17	100.00	191.16	204.53	206.35	100.00	206.21
$y(2)$-Quadratic	RB	-1.50	0.99	-11.72	0.90	-1.57	-0.52	-7.26	-0.57	-0.76	-0.12	-4.63	-0.16
	CV	36.40	41.04	33.60	40.86	25.65	26.92	25.25	26.92	20.27	20.87	20.27	20.85
	RRMSE	102.38	115.35	100.00	114.84	97.82	102.51	100.00	102.50	97.55	100.36	100.00	100.26
$y(3)$-Exponential	RB	-7.25	2.71	-32.99	2.30	-6.05	-2.03	-21.94	-2.04	-2.77	-0.28	-14.13	-0.38
	CV	135.05	155.62	104.99	154.76	94.03	99.74	81.75	99.68	75.69	78.41	68.30	78.29
	RRMSE	122.90	141.43	100.00	140.65	111.32	117.87	100.00	117.80	108.59	112.41	100.00	112.25
$y(4)$-Bump	RB	-2.81	1.29	-19.40	1.16	-2.51	-0.69	-11.91	-0.78	-1.26	-0.20	-7.63	-0.23
	CV	61.30	67.64	58.34	67.39	44.84	46.64	44.49	46.63	35.08	35.92	35.10	35.89
	RRMSE	99.81	110.04	100.00	109.62	97.50	101.27	100.00	101.26	97.71	100.00	100.00	99.90

TABLE 7.8 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 5\%$ and balancing variable x(2-Mixtured distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-4.19	-0.69	-0.21	-1.57	-2.71	-0.62	-0.28	-1.05	-0.91	0.10	0.17	-0.09
	CV	36.19	47.99	19.55	44.85	29.09	31.72	14.91	31.20	20.46	21.33	10.64	21.20
	RRMSE	186.30	245.44	100.00	229.49	195.98	212.83	100.00	209.41	192.47	200.46	100.00	199.28
$y(2)$-Quadratic	RB	-10.89	-1.05	-16.98	-3.15	-5.13	-0.49	-8.98	-1.27	-2.82	-0.12	-5.84	-0.48
	CV	44.24	82.21	35.87	74.02	35.00	44.70	30.15	42.98	29.07	33.14	25.24	32.61
	RRMSE	114.81	207.17	100.00	186.69	112.48	142.13	100.00	136.70	112.73	127.94	100.00	125.90
$y(3)$-Exponential	RB	-53.00	-8.52	-68.94	-17.64	-24.12	-2.23	-40.51	-6.16	-12.63	0.03	-27.01	-1.62
	CV	220.25	441.77	154.62	397.27	192.64	251.04	162.12	240.34	164.18	188.93	146.82	185.71
	RRMSE	133.82	261.00	100.00	234.90	116.18	150.23	100.00	143.87	110.31	126.56	100.00	124.41
$y(4)$-Bump	RB	-13.49	-1.02	-21.06	-3.60	-6.38	-0.54	-11.04	-1.48	-3.46	-0.08	-7.16	-0.51
	CV	57.02	101.73	46.52	91.99	44.44	55.88	38.03	53.83	36.71	41.51	31.57	40.88
	RRMSE	114.76	199.24	100.00	180.29	113.39	141.12	100.00	136.01	113.89	128.23	100.00	126.29

TABLE 7.9 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of Bernoulli rejectif samples with $\tau = 5\%$ and balancing variable x(3-Lognormal distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	0.24	0.34	0.24	0.85	0.02	0.03	0.11	0.30	-0.11	-0.16	-0.06	0.04
	CV	21.82	22.12	19.03	21.86	17.29	17.37	13.66	17.30	13.98	14.02	9.43	13.98
	RRMSE	114.62	116.19	100.00	114.91	126.60	127.13	100.00	126.65	148.23	148.66	100.00	148.27
$y(2)$-Quadratic	RB	-2.07	-0.23	-2.41	-1.89	-0.62	0.12	-0.96	-0.53	-0.42	-0.02	-0.73	-0.37
	CV	17.47	18.59	17.55	17.68	12.41	12.72	12.47	12.48	9.97	10.09	10.02	9.99
	RRMSE	99.31	104.96	100.00	100.34	99.36	101.69	100.00	99.84	99.27	100.38	100.00	99.52
$y(3)$-Exponential	RB	-1.30	-0.09	-1.52	-0.88	-0.42	0.09	-0.60	-0.22	-0.30	-0.05	-0.50	-0.18
	CV	13.92	15.29	12.98	14.32	10.41	10.78	9.09	10.54	8.98	9.12	7.50	9.04
	RRMSE	106.93	116.97	100.00	109.75	114.43	118.38	100.00	115.75	119.53	121.35	100.00	120.29
$y(4)$-Bump	RB	-5.43	-0.63	-6.34	-4.61	-2.12	-0.22	-2.98	-1.74	-1.33	-0.35	-2.12	-1.11
	CV	53.10	54.12	53.39	53.19	37.16	37.43	37.36	37.19	29.16	29.27	29.32	29.17
	RRMSE	99.28	100.68	100.00	99.30	99.29	99.88	100.00	99.32	99.29	99.58	100.00	99.30

TABLE 7.10 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of simple random sampling rejectif samples with $\tau = 5\%$ and balancing variable x(1-Normal distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	0.03	0.33	0.17	1.08	-0.31	-0.10	0.08	0.18	-0.37	-0.16	-0.03	-0.07
	CV	24.26	27.78	20.61	24.53	16.62	17.13	12.64	16.72	15.08	15.32	10.61	15.14
	RRMSE	117.68	134.76	100.00	119.13	131.55	135.49	100.00	132.30	142.22	144.42	100.00	142.75
$y(2)$-Quadratic	RB	-8.36	0.45	-9.24	-6.69	-3.54	0.08	-4.49	-2.62	-1.92	0.05	-2.95	-1.36
	CV	34.18	50.97	33.53	38.06	25.34	29.39	25.11	26.56	20.37	22.04	20.20	20.93
	RRMSE	101.16	146.53	100.00	111.11	100.30	115.22	100.00	104.63	100.22	107.96	100.00	102.77
$y(3)$-Exponential	RB	-26.00	1.73	-28.27	-19.12	-11.49	-0.07	-14.07	-7.81	-6.37	-0.05	-9.29	-4.10
	CV	113.52	181.45	109.36	130.16	86.40	102.61	84.25	91.62	70.45	77.23	68.22	72.87
	RRMSE	103.11	160.65	100.00	116.47	102.04	120.13	100.00	107.65	102.75	112.18	100.00	106.02
$y(4)$-Bump	RB	-13.23	0.71	-14.71	-10.31	-5.71	-0.03	-7.27	-4.16	-2.93	0.20	-4.58	-1.98
	CV	57.85	80.91	57.13	63.01	42.15	47.67	41.95	43.79	33.66	35.94	33.50	34.43
	RRMSE	100.59	137.16	100.00	108.23	99.90	111.97	100.00	103.32	99.95	106.31	100.00	102.00

TABLE 7.11 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of simple random sampling rejectif samples with $\tau = 5\%$ and balancing variable x(2-Mixtured distribution).

		$n = 25$				$n = 50$				$n = 75$			
		BEE	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.	*BEE*	MC	Fuller	Edge.
$y(1)$-Linear	RB	-0.45	0.03	-0.19	0.03	0.02	0.25	0.40	1.01	-0.21	0.21	0.10	0.56
	CV	20.59	50.36	19.08	24.05	16.16	18.39	13.74	16.60	13.90	14.84	10.92	14.24
	RRMSE	107.94	263.95	100.00	126.04	117.61	133.79	100.00	121.04	127.26	135.90	100.00	130.44
$y(2)$-Quadratic	RB	-15.05	0.18	-15.31	-10.41	-6.62	0.19	-7.00	-3.04	-3.49	0.46	-3.79	-1.23
	CV	37.45	125.09	36.66	59.48	32.59	48.86	31.43	40.17	26.91	33.52	25.91	30.32
	RRMSE	101.60	314.84	100.00	151.98	103.29	151.73	100.00	125.09	103.62	128.02	100.00	115.86
$y(3)$-Exponential	RB	-73.03	1.52	-73.79	-53.10	-31.52	2.66	-33.68	-15.37	-17.38	2.75	-19.49	-7.00
	CV	166.36	678.17	161.71	299.73	182.03	278.34	177.74	227.08	157.36	198.06	154.53	178.00
	RRMSE	102.21	381.53	100.00	171.25	102.12	153.87	100.00	125.81	101.65	127.18	100.00	114.38
$y(4)$-Bump	RB	-18.37	0.04	-18.69	-12.39	-8.06	0.18	-8.52	-3.58	-4.22	0.59	-4.59	-1.38
	CV	46.74	149.51	45.87	72.18	40.01	59.00	38.69	48.81	33.22	40.94	32.08	37.18
	RRMSE	101.39	301.85	100.00	147.86	103.01	148.90	100.00	123.53	103.32	126.35	100.00	114.81

TABLE 7.12 – Monte Carlo Relative Bias, Coefficients of variation and RRMSE (in %) of simple random sampling rejectif samples with $\tau = 5\%$ and balancing variable x(3-Lognormal distribution).

Bibliography

Breidt, F. J. and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141(1):479 – 487.

Chauvet, G. (2011). On variance estimation for the french master sample. *Journal of Official Statistics*, 27(4):651.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm for balanced sampling. *Computational Statistics*, 21(1):53–62.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91(4):893–912.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2):269–278.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4):933–944.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35:1491–1523.

Hájek, J. and Dupač, V. (1981). *Sampling from a finite population*, volume 37. M. Dekker.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Kim, J. K. and HAZIZA, D. (2013). Doubly robust inference with missing survey data. *Statistica Sinica, Accepted for publica- tion*.

Kott, P. S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. In *Survey Research Methods*, volume 6, pages 105–111.

Legg, J. C. and Yu, C. L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, pages 69–79.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337.

Royall, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, pages 657–664.

Särndal, C.-E. and Wright, R. L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11(3):pp. 146–156.

Thompson Mary, E. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34(1):3–10.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Wright, R. L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78(384):879–884.

Conclusion

Le thème central des travaux présentés dans cette thèse est la mobilisation de l'information auxiliaire en sondage. On a vu que l'utilisation de l'information auxiliaire pouvait poursuivre plusieurs objectifs : (1) produire des estimateurs qui soient cohérents avec l'information auxiliaire, (2) produire des estimateurs ajustés pour la non-réponse, (3) produire des estimateurs conditionnellement sans bais et (4) sélectionner un échantillon équilibré.

Dans le chapitre 3, on a défini un calage sur des paramètres de calage plus complexes que les totaux des variables auxiliaires. Le principe était de calculer des poids de calage pour que les données d'une enquête soient cohérentes avec des ratios ou des médianes connues par exemple. Cette nouveauté offre aux statisticiens publics une palette élargie d'estimateurs par calage qu'ils peuvent mettre en œuvre avec les outils de calage actuels.

A plusieurs reprises au cours de cette thèse, on a rappelé que si la variable d'intérêt suivait un modèle de régression linéaire en fonction des variables auxiliaires alors n'importe quel estimateur linéaire pondéré, calé sur les totaux des variables auxiliaires, était approximativement sans biais et de variance plus faible que l'estimateur Horvitz-Thompson pour des tailles d'échantillon finies. En conséquence, pour des variables d'intérêt qui suivent des modèles de régression linéaires en fonction des variables auxiliaires, il est acquis (1) qu'un estimateur par calage est adapté à la correction de la non-réponse, (2) qu'un estimateur par calage est conditionnellement sans biais et (3) que l'estimateur par la régression de Fuller associé au tirage réjectif est sans biais pour des tailles d'échantillon finies.

Lorsque la variable d'intérêt ne suit pas un modèle de régression linéaire en fonction des variables auxiliaires, on a rappelé que l'estimateur par calage était sans biais à condition que les inverses des poids de calage soient des estimateurs convergents des probabilités de sélection (conditionnelles ou marginales).

Au chapitre 4, on a étudié la correction de la non-réponse dans le cas où la variable d'intérêt suit un modèle en fonction des variables auxiliaires qui n'est pas une régression linéaire. On a montré que si la fonction de calage choisie ne correspondait pas à l'inverse de la fonction de lien dans le modèle de non-réponse alors l'estimateur par calage pouvait présenter des biais importants.

Dans le chapitre 5, on a étudié le cas particulier où la variable explicative de la

non-réponse, appelée variable instrumentale, n'est observée que sur l'échantillon des répondants. On a pris le cas simple où la variable d'intérêt suit un modèle de régression (avec constante) en fonction de la variable instrumentale. Sous le modèle de super-population de la variable d'intérêt, on a montré que l'estimateur par calage généralisé était sans biais à la condition d'utiliser des variables de calage non corrélées à l'indicatrice de réponse conditionnellement à la variable instrumentale.
On a également mis en évidence que la variance de l'estimateur par calage était inversement proportionnelle au coefficient de corrélation entre la variable instrumentale et la variable auxiliaire.
Enfin, on a établi le manque de robustesse de cette approche lorsque la variable auxiliaire choisie (ou imposée) intervient dans le modèle de réponse en plus des variables instrumentales. Le biais de l'estimateur par calage s'amplifie à la vitesse de l'inverse du coefficient de corrélation entre la variable auxiliaire et la variable instrumentale.

Dans le chapitre 6, on s'est intéressé aux estimateurs linéaires pondérés conditionnellement sans biais pour des tailles d'échantillon finies. On a proposé une méthode de calcul exacte des probabilités d'inclusion conditionnelles pour un tirage de Poisson de taille fixe et une méthode générale d'estimation des probabilités d'inclusion conditionnelles par simulation Monte Carlo. Cette méthode nécessite la connaissance des valeurs des variables auxiliaires pour toutes les unités de la base de sondage. Il s'agit d'une situation habituelle pour les enquêtes entreprises où des fichiers administratifs deviennent disponibles au moment de l'estimation. Les probabilités d'inclusion conditionnelles estimées pourraient avantageusement remplacer les probabilités d'inclusion initiales dans la procédure de calage. L'application de ce traitement aux valeurs extrêmes (sur les variables auxiliaires) et aux strata-jumpers semblent particulièrement prometteuse.

Enfin, au chapitre 7, on a examiné les propriétés de l'estimateur par la régression pondéré par les poids d'échantillonnage initiaux d'un tirage réjectif de taille finie. On a montré que les poids d'un tel estimateur s'écartaient fortement des inverses des probabilités d'inclusion réelle (ou conditionnelles au critère de rejet) lorsque la distribution de la variable d'équilibrage était asymétrique. Les simulations ont également attesté que, lorsque la variable d'intérêt ne suit pas un modèle de régression linéaire sur les variables d'équilibrage, l'estimateur par la régression (pondéré par les poids d'échantillonnage initiaux) présentait des biais importants sans être nécessairement plus efficace que l'estimateur simplement pondéré par les poids d'échantillonnage initiaux.

Bibliographie

Andersson, P. G. (2006). A conditional perspective of weighted variance estimation of the optimal regression estimator. *Journal of Statistical Planning and Inference*, 136(1) :221–234.

Andersson, P. G. and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31(1) :95–99.

Basu, D. (1971). An essay on the logical foundations of survey, part one,. In Godambe, V. P. and Sproutt, D. A., editors, *Foundations of Statistical Inference*, pages 203–233. Toronto : Holt, Rinehardt and Winston.

Berger, Y. G. (2008). A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient.

Bhattacharya, J. and Vogt, W. B. (2012). Do instrumental variables belong in propensity scores ? *International Journal of Statistics & Economics*, 9(A12) :107–127.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review / Revue Internationale de Statistique*, 51(3) :pp. 279–292.

Binder, D. A. (1991). Use of estimating functions for interval estimation from complex surveys. In *Proceedings of the survey research methods section*, pages 34–42. American Statistical Association.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples : a cookbook approach. *Survey Methodology*, 22 :17–22.

Binder, D. A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89(427) :1035–1043.

Breidt, F. J. and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141(1) :479 – 487.

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140(1) :75–91.

Chambers, R. and Clark, R. (2012). *An introduction to model-based survey sampling with applications*. OUP Oxford.

Chambers, R. L., Skinner, C., and Wang, S. (1999). Intelligent calibration. *Bulletin of the International Statistical Institute*, 58(2) :321–324.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95(3) :555–571.

Chauvet, G. (2011). On variance estimation for the french master sample. *Journal of Official Statistics*, 27(4) :651.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm for balanced sampling. *Computational Statistics*, 21(1) :53–62.

Chen, X.-H., Dempster, A. P., and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3) :457–469.

Clark, R. G. and Chambers, R. L. (2008). Adaptive calibration for prediction of finite population totals. *Survey Methodology*.

Coquet, F. and Lesage, E. (2012). Conditional inference with a complex sampling : exact computations and monte carlo estimations. *In revision*.

D'Arrigo, J. and Skinner, C. J. (2010). Linearization variance estimation for generalized raking estimators in the presence of nonresponse. *Survey methodology*, 36 :181–192.

Dell, F. and d'Haultfœuille, X. (2008). Measuring the evolution of complex indicators : Theory and application to the poverty rate in France. *Annals of Economics and Statistics / Annales d'Économie et de Statistique*, (90) :pp. 259–290.

Demnati, A. and Rao, J. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30(1) :17–26.

Deville, J.-C. (1996). Estimation de la variance du coefficient de Gini estimé par sondage. *Actes des journées de méthodologie statistique*, pages 269–288.

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. In *Actes du colloque de la Société Statistique du Canada, Sherbrooke, Canada*.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, 25 :219–230.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, France. In French.

Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des journées de méthodologie statistique*, pages 4–20.

Deville, J.-C., Sarndal, C., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, pages 1013–1020.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418) :376–382.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling : the cube method. *Biometrika*, 91(4) :893–912.

Dupacova, J. (1979). A note on rejective sampling. *Contributions to Statistics (J. Hajek memorial volume)*, Academia Prague :71Ů78.

Dupont, F. (1996). Calage et redressement de la non-réponse totale. In *Actes des journées de méthodologie statistique, 15 et 16 décembre 1993*, number 56. INSEE-Méthodes.

El Tinge, J. L. and Yansaneh, I. S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the US consumer expenditure survey. *Survey methodology*, 23 :33–40.

Estevao, V. M. and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16(4) :379–400.

Estevao, V. M. and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2) :127–147.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs : A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2) :269–278.

Fuller, W. A. (2002). Regression estimation for survey samples (with discussion). *Survey Methodology*, 28(1) :5–23.

Fuller, W. A. (2009a). *Sampling Statistics*. Wiley.

Fuller, W. A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4) :933–944.

Glasser, G. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, pages 648–654.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31 :1208–1211.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population : Their relationship and estimation. *International Statistical Review*, 54 :127–138.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35 :1491–1523.

Hájek, J. (1971). Comment on" an essay on the logical foundations of survey sampling" by d. basu. *Foundations of Statistical Inference (eds VP Godambe and DA Sprott). Toronto : Holt, Rinehart, and Winston.*

Hájek, J. and Dupač, V. (1981). *Sampling from a finite population*, volume 37. M. Dekker.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *J. Am. Stat. Assoc.*, 78 :776–793.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32 :37–52.

Haziza, D. and Lesage, E. (2013). A discussion of weighting procedures in the presence of unit nonresponse. *Submitted for publication.*

Hidiroglou, M. H., Rao, J. N. K., and Yung, W. (2002). Estimating equations for the analysis of survey data using poststratification information. *Survey Methodology*, 64(2) :364–378.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, pages 663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377) :89–96.

Kim, J. K. and HAZIZA, D. (2013). Doubly robust inference with missing survey data. *Statistica Sinica, Accepted for publica- tion.*

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35(4) :501–514.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1) :21–39.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2) :133.

Kott, P. S. (2009). Calibration weighting : Combining probability samples and linear prediction models. *Handbook of Statistics, Sample Surveys : Inference and Analysis*, 29B :55–82.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491) :1265–1275.

Kott, P. S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. In *Survey Research Methods*, volume 6, pages 105–111.

Kovacevic, M. and Binder, D. (1997). Variance estimation for measures of income inequality and polarization-the estimating equations approach. *Journal of Official Statistics*, 13(1) :41–58.

Krapavickaite, D. and Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica*, 16 :347–364.

Krieger, A. M. and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, 18(2) :225–239.

Legg, J. C. and Yu, C. L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, pages 69–79.

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2005). Does the model matter ? comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7(3) :649–673.

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29 :33–44.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24 :51–55.

Lesage, E. (2011). The use of estimating equations to perform a calibration on complex parameters. *Survey Methodology*, 37(1) :103–108.

Lesage, E. (2012). Correction de la non-réponse non ignorable par une approche modèle. In *Actes des Journées de Méthodologie Statistique de l'INSEE.*

Lesage, E. and Haziza, D. (2013). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. *Submitted for publication.*

Little, R. J. A. (1986). Survey nonresponse adjustments. *International statistical review*, 54(1) :3.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, volume 4. Wiley New York.

Little, R. J. A. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means ? *Survey Methodology*, 31(2) :161–168.

Lohr, S. L. (2009). *Sampling : design and analysis*. Thomson.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Offcial Statistics*, 15(2) :305–327.

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4) :543–570.

Montanari, G. E. (1987). Post-sampling efficient qr-prediction in large-sample surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 191–202.

Montanari, G. E. and Ranalli, M. G. (2002). Asymptotically efficient generalized regression estimators. *Journal of Official Statistics*, 18 :577–589.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11) :1213–1222.

Nascimento Silva, P. L. D. and Skinner, C. J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1) :23–32.

Osier, G. Dealing with non-ignorable non-response using generalised calibration : A simulation study based on the luxemburgish household budget survey. *Economie et Statistiques, Working papers du STATEC*, (65).

Owen, A. B. Empirical likelihood for linear models. *The Annals of Statistics*, (19) :1725–1747.

Owen, A. B. (2001). *Empirical Likelihood*. New York : Chapman and Hall.

Pearl, J. On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence Corvallis*.

Pearl, J. (2012). On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv :1203.3503*.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā : The Indian Journal of Statistics, Series B*, pages 166–186.

Plikusas, A. (2006). Non-linear calibration. In *Proceedings, Workshop on survey sampling*, Venspils, Latvia. Riga : Central Statistical Bureau of Latvia.

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11 :15–31.

Robinson, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82(399) :826–831.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2) :377–387.

Royall, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, pages 657–664.

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76(373) :66–77.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3) :581–592.

Sautory, O. (2003). Calmar 2 : une nouvelle version du programme calmar de redressement d'échantillon par calage. In *Recueil : Symposium de Statistique Canada*.

Shehzad, M. A. (2012). *Pénalisation et réduction de la dimension des variables auxiliaires en théorie des sondages*. PhD thesis, Université de Bourgogne.

Skinner, C. J., Holt, D., and Smith, T. F. (1989). *Analysis of complex surveys*. John Wiley & Sons.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2) :113–135.

Särndal, C.-E. (2011). Three factors to signal non-response bias with applications to categorical auxiliary variables. *International Statistical Review*, 79(2) :233–254.

Särndal, C.-E. and Lundström, S. (2010). Design for estimation : Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36 :131–144.

Särndal, C.-E., Lundström, S., and Wiley, J. (2005). *Estimation in surveys with nonresponse*. Wiley Hoboken, NJ.

Särndal, C.-E., Swensson, B., and Wretman., J. (1992). *Model Assisted Survey Sampling*. New-Yok : Springer-Verlag.

Särndal, C.-E. and Wright, R. L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11(3) :pp. 146–156.

Tepping, B. J. (1968). Variance estimation in complex surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pages 11–18.

Thompson, M. E. (1997). *Theory of Sample Surveys*. Chapman-Hall, London.

Thompson Mary, E. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34(1) :3–10.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities : Simple random sampling. *International Statistical Review*, 66 :303–322.

Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities : comlex design. *Survey Methodology*, 25(1) :57–66.

Tillé, Y. (2001). Théorie des sondages. *Dunod, Paris*.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). Finite population sampling and inference : a prediction approach. *Recherche*, 67 :02.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, pages 411–414.

Wooldridge, J. M. (2009). *Introductory econometrics : A modern approach*. South-Western Pub.

Wright, R. L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78(384) :879–884.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453) :185–193.

Annexe A

Calage sur les premiers axes de l'ACP

A.1 Introduction

En sondages, l'objectif du statisticien d'enquête est de choisir la stratégie qui lui permette de tirer les échantillons les plus **représentatifs** pour l'estimation de paramètres de population finie comme les totaux de plusieurs variables d'intérêt. Pour être qualifiée de représentative, la stratégie doit a minima fournir des estimations sans biais ou asymptotiquement sans biais. Lorsqu'on dispose d'information auxiliaire, on peut considérer que le fait d'estimer parfaitement des paramètres correspondant aux variables auxiliaires est un second critère de représentativité. Deux arguments étayent cette idée : d'abord il semble préférable d'avoir un échantillon qui redonne les statistiques déjà connues (argument de cohérence, conservation de la structure des données), ensuite, lorsque les variables d'intérêt sont liées linéairement aux variables auxiliaires, la précision des estimateurs est en général meilleure que celle de l'estimateur par expansion.

Les estimateurs par calage permettent d'incorporer l'information auxiliaire de façon systématique pour améliorer la représentativité de la stratégie en modifiant sa partie estimation. Dans le cas d'enquêtes disposant de nombreuses variables auxiliaires connues pour toutes les unités $k \in \mathcal{U}$, on peut être tenté d'utiliser le maximum d'information contenue dans la base de sondage : les totaux des variables auxiliaires, mais également tous leurs moments d'ordre supérieur ou les distributions, voire les corrélations entre elles... Ces calages ont déjà été proposés mais pour un nombre restreint de variables de calage. En effet, il a été montré qu'un nombre d'équations de calage trop important dégrade la précision des estimateurs par calage (Silva et Skinner, 1997). Il faut donc respecter un **principe de parcimonie** lorsqu'on utilise un estimateur par calage. Deux types de problèmes se font jour lorsque le nombre de variables de calage augmente. Le premier est lié à la résolution du système des équations de calage (voir Shehzad et al. (2012)). Si on prend l'exemple du calage linéaire, la résolution du système nécessite l'inversion de la matrice $\sum_{k \in s} d_k \boldsymbol{x}_k \boldsymbol{x}_k^T$, où $\boldsymbol{x}_k$ est le vecteur des valeurs des variables auxiliaires pour l'unité k. Il convient donc

de choisir les variables auxiliaires de telle sorte que la matrice $\sum_{k\in\mathcal{U}} \boldsymbol{x}_k\boldsymbol{x}_k^T$ soit régulière avec des valeurs propres « suffisamment grande ».
Le second problème est lié à l'augmentation de la dispersion des poids de calage qui entrainent une augmentation de la variance des estimateurs par calage (Silva et Skinner, 1997).

Comment faire, lorsqu'on dispose d'une information auxiliaire complète et pléthorique, pour choisir les variables de calage et les paramètres de calage ? Quel critère utiliser pour guider ce choix ? Une approche que nous n'examinons pas ici serait d'utiliser des modèles de superpopulation sur les variables d'intérêt pour sélectionner les variables de calage (voir Chambers et al.(1999) et Clark et Chambers (2008)). L'approche que nous présentons dans cette annexe repose sur l'analyse des données (« à la française ») et ne mobilise que les variables auxiliaires. On propose d'utiliser la méthode de l'analyse en composantes principales (ACP) pour sélectionner les variables de calage et les paramètres de calage. L'idée directrice est de considérer que l'ACP fera ressortir l'information essentielle contenue dans l'information auxiliaire.

L'objet de cette annexe est de définir la notion de calage sur une ACP et de proposer une mise en œuvre concrète de cette technique de calage.

A.2 Rappel sur l'ACP

L'information auxiliaire disponible prend la forme de p variables auxiliaires dont les vecteurs des valeurs $\boldsymbol{x}_k$ sont connus pour toutes les unités $k \in \mathcal{U}$. La base de sondage peut être représentée sous la forme d'une matrice X dont les lignes sont les vecteurs $\boldsymbol{x}_k^\top$. On note $\boldsymbol{\mu}_x$ le vecteur des moyennes des variables auxiliaires sur la population $\mathcal{U}$.

Dans une ACP, la base de sondage est vue comme un nuage de points dont les coordonnées sont les caractéristiques des individus $k \in \mathcal{U}$. Il est convenu dans cette méthode qu'une variable apporte d'autant plus d'information qu'elle étire le nuage de points (on reviendra sur ce point ultérieurment). On mesure cette étirement par la notion d'inertie qui est empruntée à la physique et qui est semblable à la notion de variance. Le principe de l'ACP consiste à procéder à une rotation du repère orthonormé initial pour offrir des premiers axes qui présentent un étirement maximal du nuage de points. La projection du nuage de points sur ces nouveaux axes donne un nouveau jeu de variables auxiliaires (les composantes principales), noté $\boldsymbol{z} = (z_1, \ldots, z_p)^\top$, qui sont rangées dans un ordre décroissant de dispersion et qui ont des covariances empiriques nulles deux à deux. On note $\boldsymbol{z}_k = (z_{k,1}, \ldots, z_{k,p})^\top$ le vecteur des valeurs des composantes principales pour l'individu k.

D'un point de vue mathématique, une ACP correspond à la diagonalisation de la

matrice de variance covariance, Σ_x, des variables auxiliaires :

$$\begin{aligned}\Sigma_x &= \frac{1}{N}(X - \boldsymbol{I}_N\boldsymbol{\mu}_x^T)^T(X - \boldsymbol{I}_N\boldsymbol{\mu}_x^T) \\ &= \frac{1}{N}X^TX - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T, \end{aligned} \tag{A.1}$$

où $\boldsymbol{I}_N$ est un vecteur unité de dimension N.
Les valeurs propres de la matrice Σ_x, notées λ_j, $j \in (1, \ldots, p)$, sont traditionnellement appelées les inerties. On les suppose toutes distinctes et strictement positives, et on les numérote par valeurs décroissantes. On note Δ la matrice diagonale des valeurs propres :

$$\Delta = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}.$$

Les vecteurs propres (normés) $\boldsymbol{u}_j$ associés aux valeurs propres λ_j sont appelés les axes principaux de l'ACP. La matrice orthogonale de changement de base P_u, qui permet de passer de la base orthonormée initiale à la base orthonormée des axes principaux, a pour lignes les vecteurs transposés $\boldsymbol{u_j}^T$.
On a

$$\Sigma_x = P_u^T \Delta P_u. \tag{A.2}$$

On obtient les composantes principales par projection du nuage de points sur les vecteurs propres :

$$\boldsymbol{z}_k = P_u\left(\boldsymbol{x}_k - \boldsymbol{\mu}_x\right). \tag{A.3}$$

On note Z la matrice dont les lignes sont les vecteurs transposés $\boldsymbol{z}_k^\top$:

$$Z = (X - \boldsymbol{I}_N\boldsymbol{\mu}_x^T)P_u^T$$

et $\Sigma_z = \dfrac{1}{N}Z^TZ$ la matrice de variance covariance des composantes principales.

Sans surprise, d'après (A.1) et (A.2), on obtient :

$$\Sigma_z = \Delta. \tag{A.4}$$

A.3 Calage sur l'ACP

Que veut dire caler sur l'ACP ? D'un point de vue théorique, cela veut dire que l'échantillon s sélectionné doit être représentatif de la forme du nuage de points de la population totale. En d'autres termes, on veut que l'ACP sur l'échantillon (pondérée par les poids de calage) fournisse les mêmes résultats que l'ACP sur la population totale $\mathcal{U}$. On aura donc une estimation exacte des vraies valeurs des valeurs propres

et des composantes principales.

D'un point de vue pratique, il s'agit d'utiliser les variables auxiliaires $\boldsymbol{x}_k$ et de caler sur les totaux et la matrice de variance covariance de ces variables auxiliaires :

$$\begin{cases} \widehat{\Sigma}_{x,CAL} = \Sigma_x \\ \hat{\boldsymbol{\mu}}_{x,CAL} = \boldsymbol{\mu}_x, \end{cases} \tag{A.5}$$

où

$$\hat{\boldsymbol{\mu}}_{x,CAL} = \frac{1}{\sum_{k\in s} w_k} \sum_{k\in s} w_k \boldsymbol{x}_k$$

et

$$\widehat{\Sigma}_{x,CAL} = \frac{1}{\sum_{k\in s} w_k} \left(\sum_{k\in s} w_k \boldsymbol{x}_k \, \boldsymbol{x}_k^T \right) - \hat{\boldsymbol{\mu}}_{x,CAL} \hat{\boldsymbol{\mu}}_{x,CAL}^T.$$

Si on ne cale pas sur les moyennes des variables auxiliaires (ou si $\sum_{k\in s} w_k \neq N$), on est dans le cas d'un calage sur paramètres complexes et on peut utiliser un calage linéarisé tel qu'il a été défini à la section (3.4) du chapitre 3.

Le système (A.5) contient $p(p+1)/2 \; + \; p = p(p+3)/2$ équations de calage. Pour 10 variables de calage cela fait 65 équations de calage ! Il paraît peu raisonnable de procéder à un calage aussi complet pour des bases de sondage qui comportent de nombreuses variables auxiliaires.

D'après (A.3) et (A.4), on peut donner une seconde écriture des équations de calage (A.5) qui porte sur les composantes principales $\boldsymbol{z}$:

$$\begin{cases} \hat{\boldsymbol{\mu}}_{z,CAL} = \boldsymbol{\mu}_z = \mathbf{0} \\ \widehat{\Sigma}_{z,CAL} = \Delta. \end{cases} \tag{A.6}$$

L'avantage du système (A.6) par rapport au système (A.5) est qu'il permet d'envisager 2 méthodes de calage « allégé »que nous allons présenter dans les sections suivantes.

Avant de traiter ces questions de parcimonie, on peut mentionner un autre avantage de l'écriture (A.6). Elle permet de s'assurer de la non-colinéarité des variables auxiliaires qui est une condition nécessaire à la résolution des équations de calage. Sans entrer dans les détails, il convient de s'assurer que les valeurs propres de la matrice de variance-covariance Σ_x sont suffisamment grandes pour garantir que la plus petite valeur propre de $\widehat{\Sigma}_{x,CAL}$ soit positive.

A.3.1 Calage sur les inerties

Notons

$$\widehat{\Sigma}_{z,CAL} = \begin{pmatrix} \hat{\sigma}^2_{z_1,CAL} & \cdots & \hat{\sigma}_{z_1 z_p,CAL} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{z_p z_1,CAL} & \cdots & \hat{\sigma}^2_{z_p,CAL} \end{pmatrix},$$

où $\hat{\sigma}_{z_i z_j,CAL} = \dfrac{1}{\sum_{k\in s} w_k} \sum_{k\in s} w_k z_{k,i} z_{k,j}$.

On peut proposer un calage « relaché »par rapport à (A.6), en ne calant que sur les termes diagonaux de la matrice $\widehat{\Sigma}_{z,CAL}$:

$$\begin{cases} \hat{\boldsymbol{\mu}}_{z,CAL} = \boldsymbol{\mu}_z = \mathbf{0} \\ \begin{pmatrix} \hat{\sigma}^2_{z_1,CAL} \\ \vdots \\ \hat{\sigma}^2_{z_p,CAL} \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_p \end{pmatrix}. \end{cases} \tag{A.7}$$

Le nombre d'équations de calage est réduit à $2p$.

Dans le cas où l'estimation de l'inertie, $\hat{\sigma}^2_{zj,CAL}$, de l'axe principal j est nulle, il convient de ne conserver les équations de calage que pour les premières composantes principales qui ne présentent pas une estimation nulle de l'inertie.

Ce type de calage devrait permettre d'obtenir des estimateurs par calage plus précis que l'estimateur par expansion pour des variables d'intérêt qui sont liées linéairement aux variables auxiliaires et aux carrés des variables auxiliaires.

A.3.2 Calage sur un sous-espace correspondant aux premières composantes principales

Pour réduire le nombre d'équations de calage tout en conservant une partie des corrélations entre variables, on peut limiter le calage aux p' premières composantes principales. Cela revient à caler sur une ACP réduite aux variables $z_1, \cdots, z_{p'}$. Les équations de calage sont alors :

$$\begin{cases} \begin{pmatrix} \hat{\mu}_{z_1,CAL} \\ \vdots \\ \hat{\mu}_{z_{p'},CAL} \end{pmatrix} = \begin{pmatrix} \mu_{z_1} \\ \vdots \\ \mu_{z_{p'}} \end{pmatrix} \\ \begin{pmatrix} \hat{\sigma}^2_{z_1,CAL} & \cdots & \hat{\sigma}_{z_1 z_{p'},CAL} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{z_{p'} z_1,CAL} & \cdots & \hat{\sigma}^2_{z_{p'},CAL} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{p'} \end{pmatrix} \end{cases} \tag{A.8}$$

Cette méthode présente un intérêt tout particulier lorsque le modèle de superpopulation entre la variable d'intérêt et les variables auxiliaires comporte des interactions entre variables ou des variables au carré.

A.4 Discussion

L'utilisation de l'ACP dans le calage pour tirer profit d'une information auxiliaire foisonnante tout en respectant un principe de parcimonie est séduisante. Toutefois, les gains en précision pour l'estimation des totaux des variables d'intérêt dépendront du traitement des variables auxiliaires initiales.
Ainsi, la pratique habituelle de normalisation des variables semble à proscrire au profit d'une multiplication des variables auxiliaires par des coefficients d'ajustement qui tiennent compte des modèles de superpopulation des variables d'intérêt. Ainsi, si la variable auxiliaire x_1 a le coefficient β_1 dans un premier modèle et le coefficient α_1 dans un second modèle, on pourrait multiplier x_1 par le coefficient $\max(|\alpha_1|, |\beta_1|)$.

Bibliographie

Chambers, R. L., Skinner, C., and Wang, S. (1999). Intelligent calibration. *Bulletin of the International Statistical Institute*, 58(2) :321–324.

Clark, R. G. and Chambers, R. L. (2008). Adaptive calibration for prediction of finite population totals. *Survey Methodology*.

Nascimento Silva, P. L. D. and Skinner, C. J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1) :23–32.

Shehzad, M. A. (2012). *Pénalisation et réduction de la dimension des variables auxiliaires en théorie des sondages*. PhD thesis, Université de Bourgogne.